View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Illinois Digital Environment for Access to Learning and Scholarship Repository

IDENTIFICATION OF POTENTIAL CONSTITUTIVE AND TISSUE-SPECIFIC PROMOTERS AND REGULATORY MOTIFS IN *GLYCINE MAX* THROUGH DATAMINING OF PUBLICLY AVAILABLE SEQUENCE DATA

BY

KATHLEEN MARY KEATING

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Crop Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:
Associate Professor Matthew E. Hudson

ABSTRACT

This is a study of gene regulation in soybean (*Glycine max)*, specifically at the level of transcription. Analysis was performed on publicly available soybean EST data, deposited in the NCBI UniGene database, combined with the soybean genome sequence. EST libraries were grouped based on the tissue they were created from. Grouped libraries were selectively filtered to find constitutive, leaf-preferential (not in seed or seed coat), and root-specific transcripts. Transcripts were annotated using *G. max* and *Arabidopsis thaliana* databases. Gene ontology term enrichment tools were used to visualize the functions of abundant transcripts for the constitutive, leaf-preferential and root-specific groups. Promoter motif analysis was performed on the promoters for themost highly expressed transcripts of each expression category. The analysis identified many probable constitutive, leaf-preferential and root-specific promoters, as well as several over-represented motifs.

ACKNOWLEDGEMENTS

I would like to thank a number of people, without whom I would not have been able to finish this project. I would like to thank my advisor, Matt, for his guidance in my research, coursework and while writing my thesis. I would also like to thank everyone in the Hudson lab, for their suggestions and many spirited discussions. Also, I would also like to thank my committee members for their helpful feedback. Further, I would like to thank Dow AgroSciences for providing me with a fellowship for my research and studies.

TABLE OF CONTENTS

CHAPTER I INTRODUCTION

The regulation of plant gene expression, specifically transcription, is a complicated process that has yet to be fully elucidated. Plant DNA, like all eukaryotes, is packaged with octamers of histones into nucleosomes, and then further into chromatin (Singh, 1998). However, histones prevent the transcriptional machinery from accessing genes (Singh, 1998; Wasserman and Sandelin, 2004). Thus, to transcribe an RNA sequence, the chromatin around the gene of interest must be remodeled (Singh, 1998; Wasserman and Sandelin, 2004). One common method of chromatin remodeling is histone acetylation (Singh, 1998). This process decreases a histone's attraction to DNA. Next, the nucleosome will likely unwind, allowing the transcriptional machinery to enter (Singh, 1998). RNA Polymerase II (Pol II), in conjunction with a number of proteins called transcription factors (TFs), bind upstream of the coding sequence of the gene. The polymerase makes a complementary RNA strand until reaching the terminator (Gibney and Nolan, 2010; Singh, 1998). Next, the RNA will undergo post-transcriptional processing (gain of 5' cap, 3' poly A tail, etc.) before being translated into a functional protein (Gibney and Nolan, 2010).

The sequence to which the RNA Pol II binds is called the promoter, specifically the core promoter. The promoter is upstream (5') of the protein coding sequence of a gene (the CDS), and plays a critical role in regulating the expression of the downstream gene (Brown, 2010). The frequency of initiation of mRNA at the promoter determines the expression level of the gene. Promoters that frequently induce transcription are designated as strong promoters (Brown, 2010). Such promoters regulate genes that must produce large numbers of transcripts within a cell. Weak promoters control genes whose products are required in the cell in smaller quantities (Brown, 2010).

Promoters include regulatory elements, or motifs, which are short sequences to which TFs bind in a sequence-specific manner to encourage or inhibit transcription (Jones and Pevzner, 2004; Tompa et al., 2005). Motifs occur most frequently in proximal promoters, but can occur in distal promoter sequences up to 10kb away from the gene they regulate (Jones and Pevzner, 2004; Kaufmann et al., 2010). Only a fraction of all extant promoters

(and their corresponding regulatory motifs) have been fully experimentally elucidated. An informatics approach is useful in motif characterization for promoters, since genes that share a common expression pattern are likely to share regulatory motifs, and these can thus be discovered by sequence similarity (Tompa et al., 2005).

A common method for *in silico* promoter characterization is to look for regulatory motifs within the promoters of a set of co-regulated genes, usually defined by an expression profiling experiment. Many useful tools exist for this purpose, with a variety of different algorithmic implementations. Unfortunately, there are many drawbacks when using computational motif discovery. Motifs are short sequences, so local alignment tools should be used (Tompa et al., 2005). If global alignments were to be used, motifs could be falsely detected. Further, motif sequences are not rigidly conserved; there can be variation in the base frequency at each position in the motif (Tompa et al., 2005). Thus, consensus sequences can be reductionist in representing motif sequences; for this reason, position matrices and position weight matrices can be used as a computational way to represent the variation between different motifs known to have the same function (Tompa et al., 2005; Stormo, 2000; Jones and Pevzner, 2004). Finally, most motif tools make the assumption that TFs bind independently, which is well known not to be the case *in vivo* (Tompa et al., 2005). When binding to promoters, TFs usually are in cassettes, forming complexes with many other proteins (Tompa et al., 2005). Further, these tools assume nucleotides within a motif do not influence each other, but this may not be the case *in vivo* (Tompa et al., 2005).

There are several approaches to discovering motifs in promoters. One common approach, "phylogenetic footprinting", incorporates phylogeny to predict TF binding site motifs (Tompa et al., 2005). Promoter regions for orthologous genes of closely related species are used to predict motifs (Tompa et al., 2005). This approach can be quite successful, for it decreases false positives. However, it requires genome information for many species (Siddharthan et al., 2005). Another disadvantage of this technique is that it requires confident determination of perfectly orthologous sets of genes, which can rarely be done without error using computational tools alone. Phylogenetic footprinting algorithms assume that sequences that are important in gene regulation are conserved, and will evolve

slower than other non-coding sequences; thus, motifs should be able to be discovered (Wasserman and Sandelin, 2004; Tompa et al., 2005). The general methodology of these algorithms is to: align promoters of orthologous genes (using programs such as Dialagn or MLAGAN), identify regions that are conserved between the promoters of these species, and score such regions (Tompa et al., 2005; Sinha et al., 2004; Siddharthan et al., 2005).

PhyloGibbs and PhyloME (Phylogenetic Motif Elicitation) are examples of phylogenetic footprinting tools (Siddharthan et al., 2005; Sinha et al., 2004). PhyMe uses expectation maximization to search a single motif at a time. This tool looks for over-represented motifs in the promoters of interest and assumes that these motifs are conserved among species (Sinha et al., 2004). Thus, PhyMe considers promoters that are assumed to be co-regulated and those that are orthologous between species. PhyloGibbs uses Gibbs sampling to look for many motifs in parallel (Siddharthan et al., 2005).

The pattern-driven or enumerative method is another common approach in algorithms for motif finding. In this approach, all possible words (using vocabulary of A, C, G, T, or N) of a specified length are enumerated. The enumerated words represent possible motifs. Frequencies for all words are calculated for the promoters of a co-regulated gene set. These word frequencies are compared with the word frequencies for the promoters in the rest of the genome. A statistical test is performed with the word frequencies for the co-regulated and reference set. When a word is significantly over-represented in the co-regulated promoter set, it is assumed to have an important biological signal (e.g. induction of translation) (Jones and Pevzner, 2004). This group of enumerative motif-finding tools are prone to both Type 1 and Type 2 error. Type 2 error is a particular problem, and both types of error increase with larger sets of co-regulated genes. Also, enumerative algorithms are prone to detect tandem repeats as overrepresented motifs. This can be corrected for computationally, but the potential for errors such as this demonstrates the need for experimental verification of computational results.

An example of a pattern-driven motif-finding tool would be Sift. It was originally published as a command line tool that detected novel motifs in *Arabidopsis thaliana* Affymetrix array

data (Hudson and Quail, 2003). Sift since has been updated with a new motif detection algorithm and is available as a web interface. Also, it can be run with *Arabidopsis* or *Glycine max* (soybean) co-regulated gene data (Walley et al., 2007, available at http://stan.cropsci.uiuc.edu/tools.php). Considering Sift uses the enumeration method to detect over-represented motifs, it evaluates millions of possible motifs for significance. Only one motif per promoter is considered, to reduce the problems caused by tandem repeats. The program calculates a statistic comparing the motif frequency between co-regulated promoters and reference promoters with the Binomial distribution. Significant motifs, their p-values, and their frequency in co-regulated and reference promoters are reported.

Elefinder is a motif discovery tool that is similar in its methodology to Sift (available at http://stan.cropsci.uiuc.edu/tools.php). Both Sift and Elefinder use a Binomial calculation to evaluate the significance of motifs present in co-regulated gene sets. However, Elefinder only searches for known and characterized plant promoter motifs, and since these rarely form tandem repeats, it considers all motifs in the promoter, increasing the statistical power of the search. This tool has previously been available for *Arabidopsis* and *G. max* promoter analysis, but has recently been updated to support 17 more plant species. Elefinder reports the significant known motifs, their p-values, and their frequency in co-regulated and reference promoters. Also, Elefinder creates graphs that display all the locations of a specific motif (for all provided co-regulated promoters) over the promoter length. This visualization is helpful in ascertaining if these promoter motifs are present in the same region in the co-regulated promoters.

For this study, we chose *Glycine max* as our model organism to study promoters and regulatory motifs. *G. max* is a major global crop, for its oil production and protein content (Schmutz et al., 2010). Within the United States, soybean accounts for 90 percent of total oilseed production. Further, 77.5 million acres of soybean were planted in 2009 in the United States, yielding $32.1 billion in farm value (USDA Economic Research Service, Soybeans and Oil Crops, http://www.ers.usda.gov/topics/crops/soybeans-oil-crops.aspx). Besides being an economically important crop, soybean is the first legume to have a

reference genome sequence available (Schmutz et al., 2010). The Williams 82 soybean genome sequence allows for studies on promoters and motifs that confer tissue-specific gene expression (Schmutz et al., 2010). One of the goals of our study was to find strong tissue-specific and constitutive promoters using an informatics approach. This methodology is quicker than single-gene cloning experiments. Important regulatory motifs involved in tissue-specific expression can be elucidated using motif discovery tools. Further investigation into transcriptional control and gene regulation mechanisms can improve crop breeding strategies. Promoters that are initially discovered and characterized *in silico* can then be experimentally isolated and cloned or synthesized. These cloned promoters can be useful in the process of gene addition (conferring expression in the tissue of interest or throughout all tissues of the plant) using transgenic means.

In order to classify promoter strength, it is possible to use the available transcript data for soybean to enumerate the number of mRNA molecules detected in different samples. Many resources were available for soybean, including various RNAseq and microarray experiments (Severin et al., 2010). However, the Soybean Expressed Sequence Tag (EST) Project had the most representative set of transcript data from large numbers of different tissues. The Soybean EST Project generated 120,000 ESTs and >50 cDNA libraries (Shoemaker et al., 2002). Within this data set, cDNA libraries for many different soybean tissues, developmental stages, and stress-inducing conditions were made (Shoemaker et al., 2002). These cDNA libraries yielded 16,298 contigs and 17,336 singletons, which composed a set of 34,264 unique gene fragments, or "unigenes" (Shoemaker et al., 2002). Data from this project was deposited in the National Center for Biotechnology Information (NCBI) UniGene database for soybean (www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=3847). The UniGene database is composed of clusters that represent a unique set of genes (Wheeler et al., 2003; Pontius et al., 2002). ESTs are placed into a UniGene cluster after linking with GenBank data for the respective organism (Wheeler et al., 2003; Pontius et al., 2002). Sequences within a UniGene cluster have extremely similar 3' untranslated regions (3' UTRs) (Wheeler et al., 2003; Pontius et al., 2002). UniGene data has previously been used to show differential

expression (up or downregulation) between tissues, but our analysis was for tissue-specific or constitutive expression.

ESTs are now an obsolete way to sequence the transcriptome, since the cost per sequence and number of sequences per sample are greatly inferior to those that can be obtained by "next generation" sequencing methods (Nagaraj et al., 2007; Hudson, 2007). However, they have the advantage of yielding sequences with lengths of hundreds of base pairs, significantly longer than "next generation" methods. Although the EST sequences are present in smaller numbers than the output from newer technologies such as RNAseq, data from the Soybean EST Project is sufficient to find strong tissue-specific or constitutive promoters, since these represent a significant percentage of total cellular mRNA.

In this work, we analyzed and filtered soybean tissue-specific EST data to discover strongly expressed potential tissue-specific (root-specific), tissue-preferential (leaf-preferential, but not expressing in seed or seed coat), and constitutive transcripts. Promoters were extracted for transcripts that fit the aforementioned expression patterns. Motif finding tools were run on each set of co-regulated promoters. This study on promoter specificity and regulatory motifs will expand the current knowledge of promoters in soybean, their conserved motifs and possibly also function.

CHAPTER II RESULTS

Gene annotation

EST data was obtained from the NCBI UniGene database for soybean (www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=3847). Transcripts within this database are clustered into "unigenes" that theoretically represent genes (Pontius et al., 2002). EST libraries were chosen for analysis only if they could accurately represent the transcriptome of one tissue (See Methods). Table 1 details the 32 selected libraries. Most tissues were represented by more than one EST library, so similar tissues were grouped for ease in filtering for tissue-specific transcripts. The library-grouping scheme is shown in Table 2. Figure 1 details the pathway that we designed to find promoters that drive expression tissue-specifically, tissue-preferentially, or constitutively.

Although transcripts within data sets in the UniGene database are assigned to clusters that represent genes, this existing annotation was not enough. Soybean unigenes were annotated using homology searches against the soybean and *Arabidopsis* databases. The BLAST-Like Alignment Tool (BLAT) was used to align the longest high-quality unigene sequence to the soybean high confidence gene coding sequences (CDS) database at high stringency (97% minimum identity) (Kent, 2002; Schmutz et al., 2010). 80.13% of the Unigenes had a soybean match using this approach. The *G. max* transcript with the highest BLAT score was chosen for unigenes with multiple hits.

For this study, soybean genome annotations for the unigene set were not found to be sufficient. At the time that this analysis was performed, *G. max* did not have reliable functional annotations for genes. *Arabidopsis* was the closest organism with reliable annotations. TBLASTX was run using the soybean high confidence CDS against the TAIR9 cDNA set with an e-value cutoff of 1e-4. TBLASTX was used in preference to a nucleotide search because codon usage varies between species, especially between such diverged species as soybean and *Arabidopsis*. Homologous genes were more likely to be found when looking for similarities between the sequences translated in all six frames.

Identification of strong, constitutive promoters by annotating potentially high-expressing constitutive transcripts

In order to identify strong, constitutive promoters for biotechnology use, the soybean EST data in public databases was mined for transcripts expressed at high levels across tissues, expressed as a percentage of total transcripts (see Methods). Since the ubiquitin promoter is widely used as a constitutive driver of biotech traits, we looked for promoters as strong or stronger than ubiquitin. When filtered for transcripts that are highly expressed in all of the grouped libraries, 48 transcripts exhibited levels above that of ubiquitin (0.023%) (Supplemental Table 1). This transcript set had two problems, indicating that not all 48 promoters were likely to be useful: (1) many of the transcripts were expressed inconsistently between tissues, despite being above ubiqitin in all of them, and (2) many genes that were present at high levels across all grouped libraries were clearly annotated as photosynthetic genes. The expression of photosynthesis-related genes in non-leaf tissues was unexpected, and may be due to experimental conditions used to generate tissue for EST experiments that do not reflect normal growing conditions (e.g. plants were grown on media in growth chambers and thus roots were exposed to light).

Non-uniform expressing transcripts and photosynthetic transcripts were removed by instituting a standard deviation-like term to control the variation around the mean abundance. The condition to retain a transcript was that the difference between its maximum abundance and its minimum abundance across the tissues (defined as a mean of all tissue samples of the same type in the database) must not be greater than double its overall mean abundance for all tissues. Transcripts would also be disqualified from further analysis if they had the same *Arabidopsis* annotation as another gene that had previously been removed. After all of the criteria were applied, 17 genes remained that have probable constitutive promoters (Table 3). Of these 17 genes, 13 have a unique *Arabidopsis* annotation.

Identification of potential high-expressing leaf-preferential transcripts

After selectively filtering for transcripts that express preferentially in the leaf and not in the seed or seed coat, 1,401 transcripts were found to be expressed above the ubiquitin level in

this tissue (Supplemental Table 2). Table 4 lists the 10 most abundant of these leaf-preferential transcripts. This leaf-preferential (not in seed or seed coat) group includes many genes that are not leaf-specific; only 460 genes were identified as abundant leaf-specific transcripts.

Identification of high-expressing root-specific transcripts

We found 219 high-expressing transcripts that occur only in the root grouped library and are at least as abundant as ubiquitin (Supplemental Table 3). Table 5 contains the 10 most abundant root-specific transcripts. Further evidence is required to show that the genes in this "root specific" group are indeed expressed only in the root and nowhere else. The EST data is relatively low in sequence depth for the root tissue samples and some other tissues. It thus lacks the depth to be able to conclude that these genes are expressed exclusively in the root. However, these transcripts and promoters can be labeled "root preferential" with confidence.

Promoter extraction

Promoters were extracted using Perl scripts and command line BLAST resources. Figure 2 details this promoter extraction methodology. A soybean genomic BLAST database was made using the soybean whole chromosomal sequence file (Gma_109.fa) and the command makeblastdb (Altschul et al., 1990; Schmutz et al., 2010). An in-house Perl script extracted the promoters of interest using the blast database via a system call to the command fastacmd (Altschul et al., 1990). Further, this Perl script calculated the proper coordinates to extract a promoter region using the annotated transcription start site (TSS), as found in the soybean GFF file (Gmax_109_gene.gff3; Schmutz et al., 2010). The extraction of all promoters of the three groups was successful (Supplemental File 1, 2, and 3). Verification that the Perl script extracted sequences upstream of a gene's TSS was done by running blastn against the *Glycine max* v1.0 database (Phytozome, http://www.phytozome.org/search.php?show=blast&method=Org_Gmax_v1.0). All promoters' sequences were shown to be accurate and in the proper location. This verification step ensures that no promoters include erroneous sequences, such as that of an adjacent gene.

We also verified that promoters contained informative sequence (i.e. do not contain unknown or masked bases, represented as N). A Perl script was run to calculate how many N bases were in promoters of the three sets. All promoters had 1 or less N bases in the constitutive group. For the top 100 leaf-preferential promoters, only two promoters had large sections of uninformative sequence (>100Ns). One of these promoters was >50% uninformative bases. The root-specific group had only one promoter that had >1 N base. Since nearly all promoters were informative, none were removed from the analysis.

Over-represented novel promoter motif discovery using Sift

Sift, a novel promoter motif discovery tool (Hudson and Quail, 2003), was run on the 17 constitutive, the top 100 root-specific and the top 100 leaf-preferential 2kb length promoters to find over-represented novel motifs. Table 6 shows the top 10 motifs for the constitutive promoter set, as found by Sift. The top hit, CGTCGNTT, had a p-value of 1.45^{-07}. This motif, however, was not declared significant at the FDR cutoff of <5%. Table 7 shows the top 10 motifs found using Sift for the leaf-preferential promoters. The top motif was GNNGAACTC, which had a p-value of 2.53^{-06}. This motif did not pass the FDR cutoff of <5%. Table 8 shows the top 10 Sift results for the root-specific promoter set. The most significant motif was GNAATNTCA, which had a p-value of 8.27^{-09} and was declared significant at FDR <5%. It is not unexpected that only one motif was shown to be significant after correcting for multiple testing error with Sift results. Sift enumerates millions of possible motifs, so it is very unlikely that a motif found will be declared significant after FDR analysis. However, since new promoter motifs reflect new regulatory pathways of gene expression and are thus routes to important scientific discoveries, higher FDR motifs can be considered. It is likely worthwhile embarking on a program of experimental validation to discover a new motif with a 90% chance of being valid, for example.

Over-represented previously-characterized promoter motif discovery with Elefinder

Elefinder is a web-based tool designed to enumerate only previously characterized motifs, and thus reduce the multiple testing problem inherent in the sift approach. It is available on the web for soybean at (available at http://www.stan.cropsci.uiuc.edu/tools.php).

Elefinder was run on minimal and 2kb promoters for the constitutive, leaf-preferential, and root-specific groups to find over-represented known promoter motifs. Table 9 shows the Elefinder results for the minimal constitutive promoters. All of the motifs found by Elefinder for this group had poor p-values (ranging from 1.11^{-1} to 2.06^{-1}), which were not declared significant at an FDR <5%. No significant motifs were found for the top 100 leaf-preferential minimal promoters. Table 10 shows the over-represented motifs for the top 100 root-specific promoters. Like the constitutive motif results, the top motifs for the root-specific promoters had poor p-values (1.99^{-1} to 3.03^{-1}) and were not significant at FDR <5%. Elefinder was not able to find significant over-represented motifs with the minimal constitutive, leaf-preferential, or root-specific promoters. This may indicate a lack of knowledge of the motifs necessary for these core promoters to function in the current literature.

Elefinder discovered more significant motifs with the 2kb promoters. Table 11 summarizes the motifs found for the constitutive promoters. No significant motifs were found after correcting for multi-testing error. Table 12 details the motifs found for the top 100 leaf-preferential 2kb promoters. Two motifs, Bellringer and DPBF1&2, were found to be over-represented and significant at FDR <5% (Bao et al., 2004; Kim et al., 1997). Figure 3b shows that both Bellringer and DPBF1&2 are clearly over-represented in the leaf-preferential promoters as compared to the reference promoters. Figure 4 shows the distribution of the top hit motif, Bellringer, among the promoter sequences. Figure 5 shows the distributions of the 2nd most significant motif, DPBF1&2, along the promoter length. Both figures show many occurrences of these motifs along the 2kb promoter sequence. However, there are sites where the position of certain motifs appear to be conserved between the leaf-preferential promoters. For example, there are many DPBF 1& 2 motif sites directly upstream of the TSS for these promoters. The root-specific promoters had 3 significantly over-represented motifs, W-box, Bellringer, and RAV1-A (Yu et al., 2011; Bao et al., 2004; Yu et al., 2001). Both W-box and Bellringer are significant at FDR < 0.0001. The top 10 motifs found by Elefinder for the root-specific promoters are found in Table 13. Figure 3a shows the over-represented motifs for the root-specific promoters. The distribution of the W-box motif can be seen in Figure 6. W-box is extremely prevalent in the

root-specific promoters, but there are several sites in which W-box appears to be conserved (e.g. around -1kb). Figure 7 shows the distribution of Bellringer. Bellringer is less abundant than W-box, but this motif occurs more often in leaf-preferential promoters than the background set (Figure 3a). Elefinder was more successful in finding significant motifs with longer promoters, which could be due to incorrectly annotated gene structure (e.g. wrongly annotated TSS).

Functional annotation visualization

REViGO was used for visualization of GO terms for each of the three categories. It was run using Arabidopsis as a reference of GO term abundance. REViGO generated R scripts to plot the results for constitutive, leaf-preferential and root-specific (shown in Figures 8, 9, 10, respectively). Interpretation of the REViGO plots is not intuitive. The plot axes (semantic space x and y) do not have an inherent meaning; they represent multi-dimensional scaling of the GO data. These plots should be interpreted by the clustering of the GO terms, their colors (representing the $\log_{10}$p-value), and their size (larger size means more abundant, which usually means more general, GO term).

Figure 8 shows the REViGO results for the constitutive promoters. The most significant GO term is "translation" for the constitutive gene set (p-value of 0.00014). Other enriched GO terms include "macromolecule metabolic process" (p-value of 0.037), "cellular protein metabolic process" (p-value of 0.0039) and "cytosol" (p-value of 0.0043). Less significant GO terms include "regulation of cellular process", "biological regulation", and "metabolic process" (p-values of 0.45, 0.24, and 0.28 respectively).

GO term enrichment was performed for the top 100 Leaf-preferential genes found using REViGO. As Figure 9 shows, the most significant GO terms for the leaf-preferential set are "photosynthesis" and "photosynthesis light reaction" (p-values of 2.4e-12 and 2.2e-05, respectively). This group contains many more GO terms than the constitutive group's REViGO plot. However, a majority of these GO terms do not have significant p-values. Of the top 25 most significant GO terms for the leaf-preferential group, 19 are related to

photosynthesis, which is consistent with expectations of the types of genes likely to be highly expressed in leaves.

REViGO was used to analyze GO term enrichment for the top 100 root-specific genes. Figure 10 shows the REViGO plot. Among the most significant GO terms for the root-specific set are “localization” and “vesicle-mediated transport” (p-values of 0.011 and 0.058, respectively). Interestingly, the leaf-preferential group and the root-specific group contain a similar number of GO terms, but the root-specific group has greater overlap between GO terms (represented as overlap on the graph) and more significant p-values.

Tables

Table 1. EST libraries used for analysis. Data shown is from build #39 of the Unigene database.

dbEST ID	Library Name	Tissue Type	Cultivar	Number of ESTs
1617	Gm-c1004	Root, 8 day old seedlings	Williams	7600
1847	Gm-c1007	Cotyledon, seedlings with individual seed fresh weight of 100-300mg, greenhouse grown	Williams	2255
1848	Gm-c1008	Pod, 2 cm long, 12-week old plants, greenhouse grown	Williams	1700
1849	Gm-c1009	Root, 2-month old plants, greenhouse grown	Williams	2329
1867	Gm-c1016	Immature flowers, field grown	Williams 82	7880
1892	Gm-c1019	Seed coat, immature (200-300mgs), greenhouse grown	Williams	4855
1957	Gm-c1014	Leaf, seedling, 2 to 3-week old, greenhouse grown	Williams	2076
2224	Gm-c1015	Mature flowers, field grown	Williams 82	4880
2233	Gm-c1018	Leaf, 2 to 3-week old, greenhouse grown	Williams 82	1229
2567	Gm-c1023	Seed coat, immature (100-200mgs)	T157	3192
2837	Gm-c1025	Hypocotyl, 3-day old seedlings	Williams 82	1055
3696	Gm-c1026	Leaf, senescing, mature plants, greenhouse grown	Williams	2179
3798	Gm-c1032	Cotyledon, seedling, 8-day old, etiolated for 3 days, greenhouse grown 3 days	Williams	2349
3799	Gm-c1033	Root, seedling	Desloy 5710	1223
4067	Gm-c1035	Leaf, seedling, immature (unfurled trifolate), greenhouse grown	Williams	1670
4110	Gm-c1036	Somatic embryos (2-9 months) cultured on MSD 20	Jack	9557
4133	Gm-c-1037	Leaf, seedling, 2-week old, greenhouse grown	Williams	1749
5364	Gm-c1040	Germinating seeds, hypocotyl &plumule	Williams 82	2945

Table 1 (cont.)

dbEST ID	Library Name	Tissue Type	Cultivar	Number of ESTs
5425	Gm-c1043	Germinating seeds, hypocotyl &plumule	Williams	3211
5568	Gm-c1046	Seeds, germinated for 3 days	Williams	1027
6110	Gm-c1051	Floral meristem	Corolla	5868
6112	Gm-c1061	Mature flowers, field grown	Raiden	3766
6113	Gm-c1062	Stem, 1-month old, stem	Raiden	5184
6114	Gm-c1047	Leaf, immature, unfurled trifolate, greenhouse grown	Williams	1510
6784	Gm-c1064	Epicotyl, 2-week old seedlings	Williams	2678
6820	Gm-c1054	Leaf, 3-week old, greenhouse grown	Harosoy	5009
7250	Gm-c1055	Pod, mature prior to senescence, greenhouse grown	L82-2024	3281
8606	Gm-c1063	Germinating shoots, 24-hour germination	Williams	3798
8651	Gm-c1071	Pod, 2 cm long, greenhouse grown	Williams	3642
8803	Gm-c1075	Somatic embryos cultured on MSM6AC	Jack	3492
9659	Gm-c1077	Cotyledon, seedlings, 18-days old, individual seed fresh weight of 100-300mg, greenhouse grown	Williams	1643
9967	Gm-c1086	Seeds, young (< 20mgs)	Williams 82	1321

Table 2. Grouped EST libraries. Libraries were grouped according to tissue type and developmental stage.

Group	dbEST ID	Library name	Tissue Type	Total Number of ESTs
Cotyledon	1847	Gm-c1007	Cotyledon, seedlings with individual seed fresh weight of 100-300mg, greenhouse grown	6247
	3798	Gm-c1032	Cotyledon, seedling, 8-day old, etiolated for 3 days, greenhouse grown 3 days	
	9659	Gm-c1077	Cotyledon, seedlings, 18-days old, individual seed fresh weight of 100-300mg, greenhouse grown	
Flower	6110	Gm-c1051	Floral meristem	22394
	1867	Gm-c1016	Immature flowers, field grown	
	2224	Gm-c1015	Mature flowers, field grown	
	6112	Gm-c1061	Mature flowers, field grown	
Leaf	6820	Gm-c1054	Leaf, 3-week old, greenhouse grown	15422
	1957	Gm-c1014	Leaf, seedling, 2 to 3-week old, greenhouse grown	
	3696	Gm-c1026	Leaf, senescing, mature plants, greenhouse grown	
	4133	Gm-c-1037	Leaf, seedling, 2-week old, greenhouse grown	
	4067	Gm-c1035	Leaf, seedling, immature (unfurled trifolate), greenhouse grown	
	6114	Gm-c1047	Leaf, immature, unfurled trifolate, greenhouse grown	
	2233	Gm-c1018	Leaf, 2 to 3-week old, greenhouse grown	

Table 2 (cont.)

Group	dbEST ID	Library name	Tissue Type	Total Number of ESTs
Pod	9967	Gm-c1086	Seeds, young (< 20mgs)	9944
	1848	Gm-c1008	Pod, 2 cm long, 12-week old plants, greenhouse grown	
	8651	Gm-c1071	Pod, 2 cm long, greenhouse grown	
	7250	Gm-c1055	Pod, mature prior to senescence, greenhouse grown	
Root	1617	Gm-c1004	Root, 8 day old seedlings	11152
	3799	Gm-c1033	Root, seedling	
	1849	Gm-c1009	Root, 2-month old plants, greenhouse grown	
Seed Coat	1892	Gm-c1019	Seed coat, immature (200-300mgs), greenhouse grown	8047
	2567	Gm-c1023	Seed coat, immature (100-200mgs)	
Somatic embryo	4110	Gm-c1036	Somatic embryos (2-9 months) cultured on MSD 20	13049
	8803	Gm-c1075	Somatic embryos cultured on MSM6AC	
Stems	6113	Gm-c1062	Stem, 1-month old, stem	18871
	8606	Gm-c1063	Germinating shoots, 24-hour germination	
	5425	Gm-c1043	Germinating seeds, hypocotyl &plumule	
	5364	Gm-c1040	Germinating seeds, hypocotyl &plumule	
	2837	Gm-c1025	Hypocotyl, 3-day old seedlings	
	6784	Gm-c1064	Epicotyl, 2-week old seedlings	
Germinating seeds	5568	Gm-c1046	Seeds, germinated for 3 days	1027

Table 3. Highly expressed constitutive transcripts identified from analysis of the grouped Unigene library data.

Overall Mean Percent Expression	Unigene ID	*Glycine max* transcript ID	*Arabidopsis thaliana* transcript ID	TAIR 9 annotation
0.316%	Gma.30081	Glyma13g17830.1	AT4G02890.4	UBQ14 | UBQ14; protein binding
0.136%	Gma.16708	Glyma11g37970.1	AT5G10980.1	histone H3
0.110%	Gma.30052	Glyma07g39020.1	AT4G21960.1	PRXR1 | PRXR1; electron carrier/ heme binding / peroxidase
0.097%	Gma.2876	Glyma02g00810.1	AT5G43330.1	malate dehydrogenase, cytosolic, putative
0.090%	Gma.1254	Glyma11g10480.1	AT2G16600.1	ROC3 | ROC3; peptidyl-prolylcis-trans isomerase
0.086%	Gma.54656	Glyma05g00780.1	AT1G69410.1	ATELF5A-3, ELF5A-3 | ELF5A-3 (EUKARYOTIC ELONGATION FACTOR 5A-3); translation initiation factor
0.085%	Gma.30809	Glyma15g13140.1	AT5G59890.1	ADF4, ATADF4 | ADF4 (ACTIN DEPOLYMERIZING FACTOR 4); actin binding
0.081%	Gma.4334	Glyma06g04840.1	AT1G27730.1	STZ, ZAT10 | STZ (salt tolerance zinc finger); nucleic acid binding / transcription factor/ transcription repressor/ zinc ion binding
0.074%	Gma.3789	Glyma04g32950.1	AT1G69410.1	ATELF5A-3, ELF5A-3 | ELF5A-3 (EUKARYOTIC ELONGATION FACTOR 5A-3); translation initiation factor
0.072%	Gma.18269	Glyma04g41750.1	AT5G53300.1	UBC10 | UBC10 (ubiquitin-conjugating enzyme 10); ubiquitin-protein ligase
0.070%	Gma.54791	Glyma10g29170.1	AT3G05590.1	RPL18 | RPL18 (RIBOSOMAL PROTEIN L18); structural constituent of ribosome
0.070%	Gma.54637	Glyma03g39420.1	AT3G05590.1	RPL18 | RPL18 (RIBOSOMAL PROTEIN L18); structural constituent of ribosome
0.068%	Gma.54647	Glyma15g42620.1	AT1G18540.1	60S ribosomal protein L6 (RPL6A)

Table 3 (cont.)

0.066%	Gma.13248	Glyma14g38620.1	AT5G41700.1	UBC8, ATUBC8 | UBC8 (UBIQUITIN CONJUGATING ENZYME 8); protein binding / ubiquitin-protein ligase
0.065%	Gma.32534	Glyma06g16050.1	AT5G64030.1	dehydration-responsive protein-related
0.056%	Gma.54790	Glyma16g23730.1	AT5G07090.1	40S ribosomal protein S4 (RPS4B)
0.053%	Gma.11306	Glyma09g38590.1	AT3G02790.1	zinc finger (C2H2 type) family protein

Table 4. Top 10 highly expressed leaf-preferential transcripts identified from analysis of the grouped Unigene library data.

Overall Mean Percent Expression	**Standard Error of Mean % Expression**	**Unigene ID**	***Glycine max* transcript ID**	***Arabidopsis thaliana* transcript ID**	**TAIR 9 annotation**
0.8468%	0.007184554	Gma.55208	Glyma11g34230.1	AT2G39730.1	RCA | RCA (RUBISCO ACTIVASE); ADP binding / ATP binding / enzyme regulator/ ribulose-1,5-bisphosphate carboxylase/oxygenase activator
0.6028%	0.001943702	Gma.32369	Glyma10g39740.1	AT5G54770.1	THI1, TZ, THI4 | THI1; protein homodimerization
0.5413%	0.004408379	Gma.12713	Glyma11g03230.1	AT1G12360.1	KEU | KEU (keule); protein transporter
0.4335%	N/A	Gma.3256	Glyma02g03980.1	AT1G23750.1	DNA-binding protein-related
0.4097%	0.00097521	Gma.30703	Glyma07g01730.1	AT4G25150.1	acid phosphatase, putative
0.3255%	N/A	Gma.32726	Glyma18g50950.1	AT5G38895.1	zinc finger (C3HC4-type RING finger) family protein
0.3101%	0.001295203	Gma.55210	Glyma19g01050.1	AT3G01500.1	CA1, ATBCA1, SABP3, ATSABP3 | CA1 (CARBONIC ANHYDRASE 1); carbonate dehydratase/ zinc ion binding
0.2890%	N/A	Gma.1621	Glyma14g05890.1	AT3G56200.1	amino acid transporter family protein
0.2876%	0.00064175	Gma.10843	Glyma13g37880.1	AT4G03280.2	PETC, PGR1 | PETC (PHOTOSYNTHETIC ELECTRON TRANSFER C); electron transporter, transferring electrons from cytochrome b6/f complex of photosystem II
0.2631%	0.000590347	Gma.31628	Glyma05g31450.1	AT5G10170.1	ATMIPS3, MIPS3 | MIPS3 (MYO-INOSITOL-1-PHOSTPATE SYNTHASE 3); binding / catalytic/ inositol-3-phosphate synthase

Table 5. Top 10 highly expressed root-specific transcripts identified from the analysis of the grouped Unigene library data.

Overall Mean Percent Expression	**Standard Error of Mean % Expression**	**Unigene ID**	***Glycine max* transcript ID**	***Arabidopsis thaliana* transcript ID**	**TAIR 9 annotation**
0.215%	N/A	Gma.6302	Glyma12g05770.1	AT2G44480.1	BGLU17 | BGLU17 (BETA GLUCOSIDASE 17); catalytic/ cation binding / hydrolase, hydrolyzing O-glycosyl compounds
0.185%	0.001420639	Gma.55058	Glyma15g13550.1	AT2G38380.1	peroxidase 22 (PER22) (P22) (PRXEA) / basic peroxidase E
0.164%	N/A	Gma.19087	Glyma05g27760.1	AT3G01750.1	ankyrin repeat family protein
0.164%	N/A	Gma.36381	Glyma04g06230.1	AT2G24762.1	AtGDU4 | AtGDU4 (Arabidopsis thaliana GLUTAMINE DUMPER 4)
0.088%	0.000751872	Gma.52820	Glyma20g22180.1	AT1G09560.1	GLP5 | GLP5 (GERMIN-LIKE PROTEIN 5); manganese ion binding / nutrient reservoir
0.088%	0.000751872	Gma.2605	Glyma07g09560.1	AT5G67488.1	other RNA
0.086%	N/A	Gma.43924	Glyma11g09060.1	AT2G17220.1	protein kinase, putative
0.086%	N/A	Gma.25544	Glyma13g01120.1	AT4G25810.1	XTR6, XTH23 | XTR6 (XYLOGLUCAN ENDOTRANSGLYCOSYLASE 6); hydrolase, acting on glycosyl bonds / hydrolase, hydrolyzing O-glycosyl compounds / xyloglucan:xyloglucosyltransferase
0.086%	N/A	Gma.6645	Glyma0041s00240.1	AT5G09530.1	hydroxyproline-rich glycoprotein family protein
0.086%	N/A	Gma.16945	Glyma05g03560.1	AT1G12610.1	DDF1 | DDF1 (DWARF AND DELAYED FLOWERING 1); DNA binding / sequence-specific DNA binding / transcription factor

Table 6. Top 10 Sift results for the 17 constitutive promoters, as ranked by increasing p-value. Sift was run with 2kb promoters.

Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR level of <5%?
CGTCGNTT	0.41	0.03	1.45076E-07	No, significant with FDR of < 10%
ANACCCNG	0.59	0.10	1.05648E-06	No
TCGGGCTAA	0.18	0.00	1.45015E-06	No
CGTCGTTT	0.29	0.01	3.38199E-06	No
GTGGANCGA	0.24	0.01	4.07625E-06	No
CGTCGNT	0.47	0.07	8.20018E-06	No
TTGTNACGC	0.24	0.01	1.06539E-05	No
TTTNGTNAC	0.71	0.21	1.21863E-05	No
CGTCGTT	0.35	0.03	1.51602E-05	No
TNACGNNG	0.88	0.37	1.53585E-05	No

Table 7. Top 10 Sift results for the top 100 leaf-preferential promoters, as ranked by increasing p-value. Sift was run with promoters with a length of 2000bp.

Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR level of 5%?
GNNGAACTC	0.16	0.04	2.52818E-06	No
ACGTGAATC	0.06	0.00	3.38745E-06	No
AATCAAAT	0.45	0.25	5.25017E-06	No
GANNTTACC	0.18	0.06	9.24849E-06	No
ATGTNCCGT	0.07	0.01	9.48375E-06	No
ACGNGAATC	0.08	0.01	9.61367E-06	No
GNTGTCTNA	0.2	0.07	1.27889E-05	No
CCNTACGNC	0.08	0.01	1.31595E-05	No
TGNAGGCC	0.13	0.03	1.56879E-05	No
GNTGAACTC	0.09	0.01	1.65004E-05	No

Table 8. Top 10 Sift results for the top 100 root-specific promoters, as ranked by increasing p-value. Sift was run with promoters with a length of 2000bp.

Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR level of 5%?
GNAATNTCA	0.43	0.18	8.27169E-09	Yes
TATTATATA	0.36	0.16	5.49463E-07	No
ANACNGTA	0.48	0.26	1.44289E-06	No
CNCCNAAT	0.53	0.31	1.66792E-06	No
GAATGTNNT	0.37	0.18	1.79336E-06	No
TATTATAT	0.57	0.34	1.80596E-06	No
TTNCNNTCG	0.35	0.16	2.93534E-06	No
TATNTGNT	0.92	0.74	2.94693E-06	No
ACCCAATG	0.11	0.02	3.90132E-06	No
CACCCAATG	0.06	0.00	4.54826E-06	No

Table 9. Elefinder results for the 17 constitutive promoters (length of 500bp).

Motif Name	Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR level of 5%?
ABFs binding site motif	CACGTGGC	0.12	0.04	1.11e-01	No
SORLIP2	GGGCC	0.59	0.56	1.89e-01	No
G-box promoter motif	CACGTG	0.29	0.27	2.01e-01	No
GATA promoter motif	(A/T)GATA(G/A)	0.12	0.06	2.06e-01	No

Table 10. Elefinder results for the top 100 root-specific promoters (length of 500bp).

Motif Name	Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR level of 5%?
ABFs binding site motif	CACGTGGC	0.04	0.038	1.99e-01	No
CBF2 binding site motif, GBF1/2/3 BS in ADH1	CCACGTGG	0.02	0.018	2.72e-01	No
SORLREP5	TTGCATGACT	0.01	0.005	3.03e-01	No

Table 11. Top 10 Elefinder results for the 17 constitutive promoters (length of 2000bp).

Motif Name	Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR level of 5%?
SORLIP2	GGGCC	1.058823529	0.56164354	1.33e-03	No
RAV1-A binding site motif	CAACA	6.235294118	5.636677529	4.83e-03	No
DPBF1&2 binding site motif	ACACNNG	1.176470588	0.744729062	5.38e-03	No
Bellringer/replumless/pennywise BS1 IN AG	AAATTAAA	2.411764706	1.925937116	5.67e-03	No
Bellringer/replumless/pennywise BS3 IN AG	ACTAATTT	0.705882353	0.395714193	7.35e-03	No
Bellringer/replumless/pennywise BS2 IN AG	AAATTAGT	0.705882353	0.395714193	7.35e-03	No
Evening Element promoter motif	AAAATATCT	0.352941176	0.162325471	3.23e-02	No
JASE1 motif in OPR1	CGTCAATGAA	0.058823529	0.002114849	3.48e-02	No
G-box promoter motif	CACGTG	0.470588235	0.265931505	3.76e-02	No
RAV1-B binding site motif	CACCTG	0.470588235	0.280951251	4.85e-02	No

Table 12. Top 10 Elefinder results for the top 100 leaf-preferential promoters (length of 2000bp).

Motif Name	Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR level of 5%?
Bellringer/replumless/pennywise BS1 IN AG	AAATTAAA	2.36	1.93	2.03e-08	Yes
DPBF1&2 binding site motif	ACACNNG	0.92	0.74	5.62e-06	Yes
SORLIP5	GAGTGAG	0.37	0.27	7.95e-03	No
Box II promoter motif	GGTTAA	1.14	1.02	1.27e-02	No
G-box promoter motif	CACGTG	0.35	0.27	1.51e-02	No
ATB2/AtbZIP53/AtbZIP44/GBF5 BS in ProDH	ACTCAT	1.34	1.24	2.03e-02	No
Z-box promoter motif	ATACGTGT	0.08	0.04	3.10e-02	No
EveningElement promoter motif	AAAATATCT	0.21	0.16	4.48e-02	No
ARF binding site motif, ARF1 binding site motif	TGTCTC	0.85	0.80	4.51e-02	No
ABFs binding site motif	CACGTGGC	0.07	0.04	5.15e-02	No

Table 13. Top 10 Elefinder results for the top 100 root-specific promoters (length of 2000bp).

Motif Name	Motif Sequence	Mean number of motifs per promoter in co-regulated set	Mean number of motifs per promoter in reference set	P-value	Significant at FDR cutoff level of 5%?
W-box promoter motif	TTGAC	4.37	3.71	2.25e-13	Yes
Bellringer/replumless/ pennywise BS1 IN AG	AAATTAAA	2.29	1.93	1.90e-06	Yes
RAV1-A binding site motif	CAACA	5.8	5.64	8.84e-04	Yes
Box II promoter motif	GGTTAA	1.18	1.02	4.07e-03	No
DPBF1&2 binding site motif	ACACNNG	0.83	0.74	1.31e-02	No
Bellringer/replumless/ pennywise BS3 IN AG	ACTAATTT	0.48	0.40	1.86e-02	No
Bellringer/replumless/ pennywise BS2 IN AG	AAATTAGT	0.48	0.40	1.86e-02	No
ATB2/AtbZIP53/AtbZIP44/GBF5 BS in ProDH	ACTCAT	1.33	1.24	2.49e-02	No
SORLREP4	CTCCTAATT	0.06	0.03	4.01e-02	No
SORLIP2	GGGCC	0.62	0.56	4.07e-02	No

Figures

Figure 1. An overview of candidate promoter identification. Applicable soybean EST libraries from the NCBI database Unigene (build #39) were chosen and used for promoter identification. Each library has the NCBI identifier (four digit number) noted. Each eligible library had its transcript count converted to a percentage for ease of comparison between libraries. Libraries from similar tissues and developmental stage were grouped. If a Unigene occurred more than once in a group (more than one library contained this), their transcript percent abundances for this Unigene were used to calculate the mean percent abundance, and the standard error.

BLAT was performed on the unigene longest high quality sequence file against *Glycine max* high confidence coding sequence file to get further annotations. Once each unigene had a soybean transcript identifier, TBLASTX was performed on the *Glycine max* high confidence coding sequence file against the TAIR9 cDNA file for better annotations from *Arabidopsis thaliana.*

Selected filtering of these datasets led to lists of candidate promoters for constitutive, leaf-preferential not in seed or seed coat, and root specific promoters. Transgenic plants will be used to validate that the promoter sets confer expression as predicted.

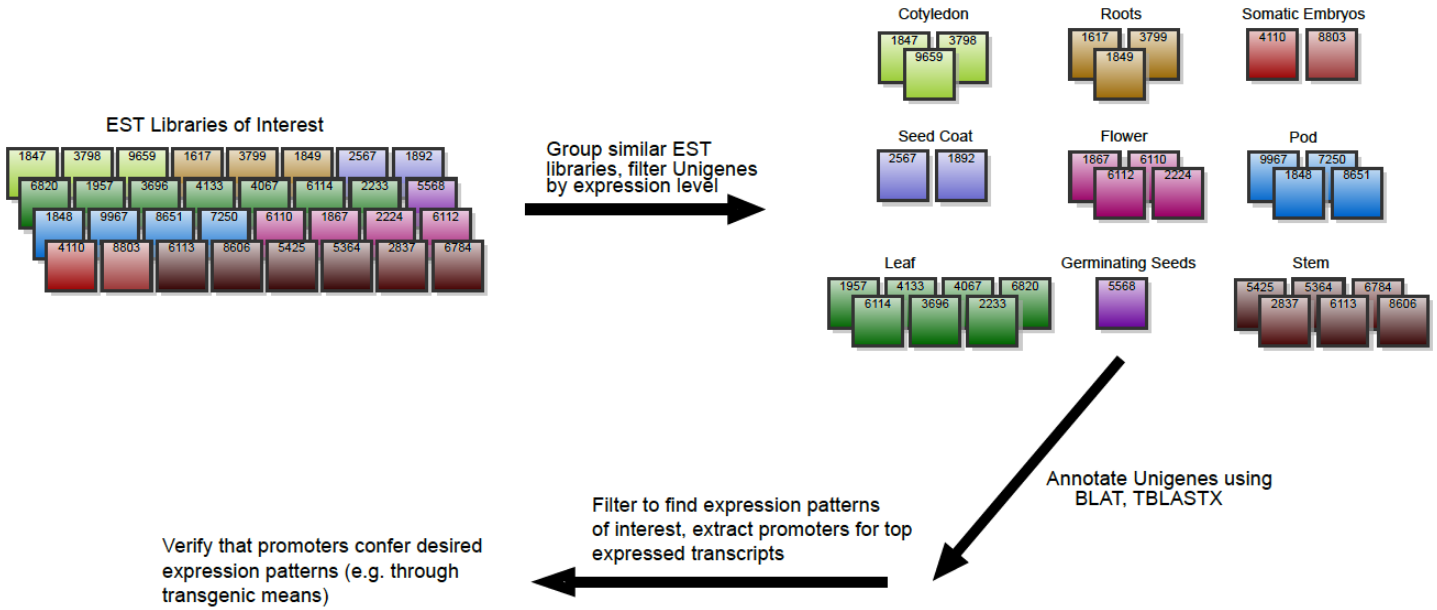


Figure 2. Promoter extraction method for genes with top expressed transcripts. Promoters were extracted by taking sequence immediately upstream of the 5' UTR, to 500bp upstream (for minimal promoter), 2kb upstream (for full promoter) or until the adjacent gene was encountered. For motif finding purposes, this method was used. Sequences were extracted using a BLAST database made from the *Glycine max* genome and a *G. max* GFF file from the Soybean Genome Project Glyma data version 1.0.

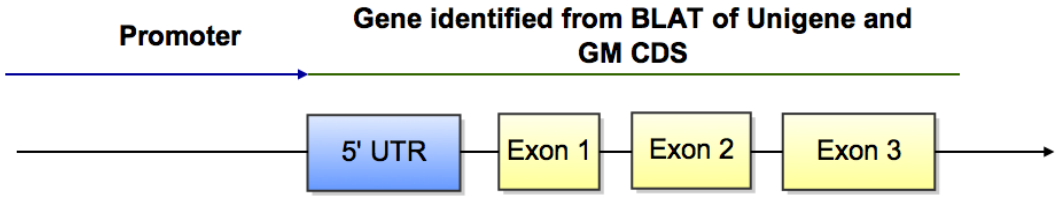
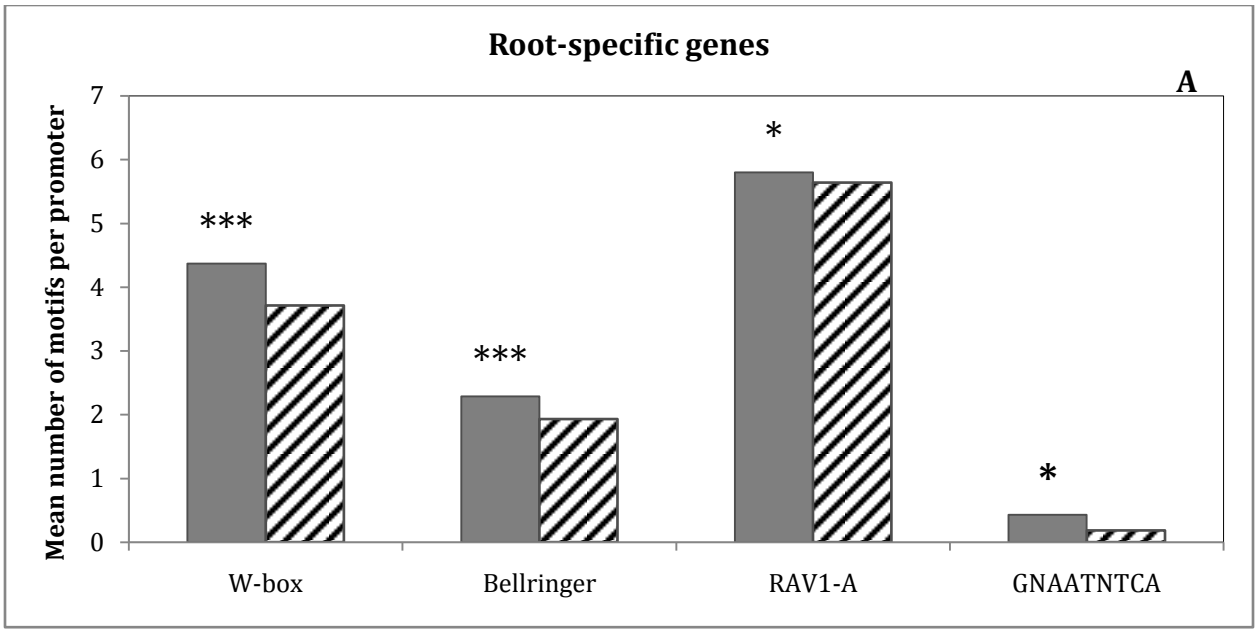


Figure 3. Over-representation of several cis-regulatory promoter motifs in co-regulated promoter sets (leaf-preferential, root-specific) found using Elefinder and Sift. (A) shows over-represented motifs in the root-specific gene set, **(B)** shows over-represented motifs in the leaf-preferential gene set. For both graphs, the mean number of motifs per promoter for the co-regulated genes are in gray, and the mean number of motifs of per promoter in the reference set are hatched. Motifs designated with names are from Elefinder; motifs labeled with a sequence are from Sift. Motif analysis was run on 2kb length promoters. * = FDR <0.05; ** = FDR <0.001; *** = FDR <0.0001.



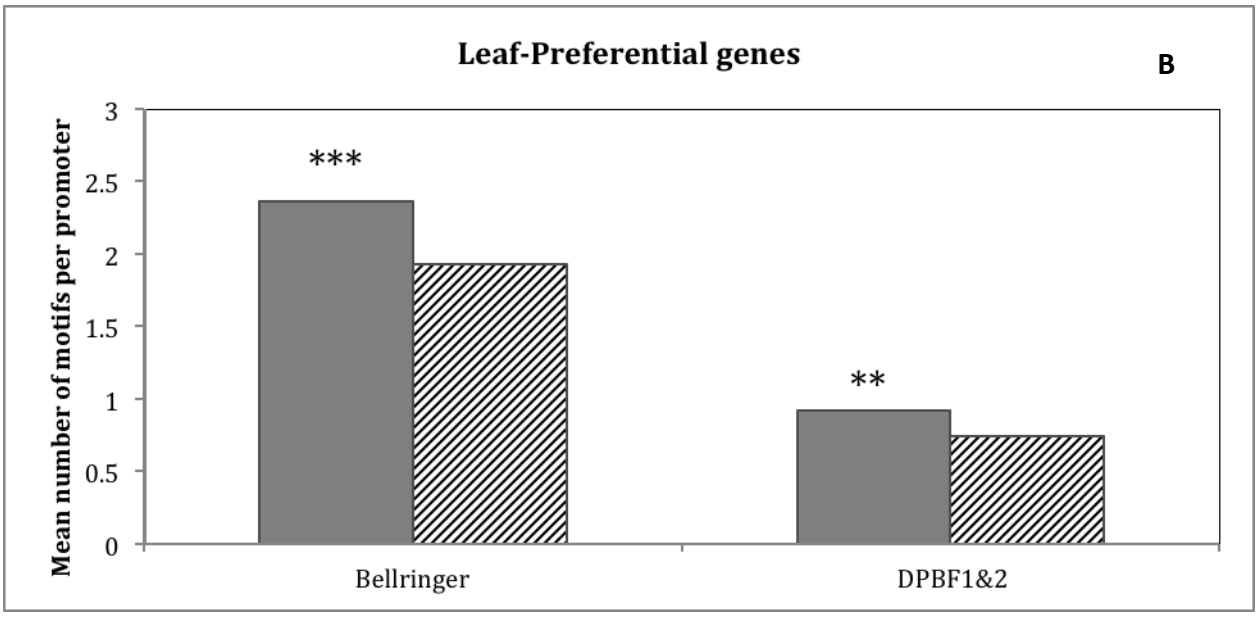


Figure 4. Distribution of the top hit motif, Bellringer/replumless/pennywise BS1 IN AG, in the promoters of the Leaf Preferential group (promoter length of 2kb). Figure was generated by Elefinder.

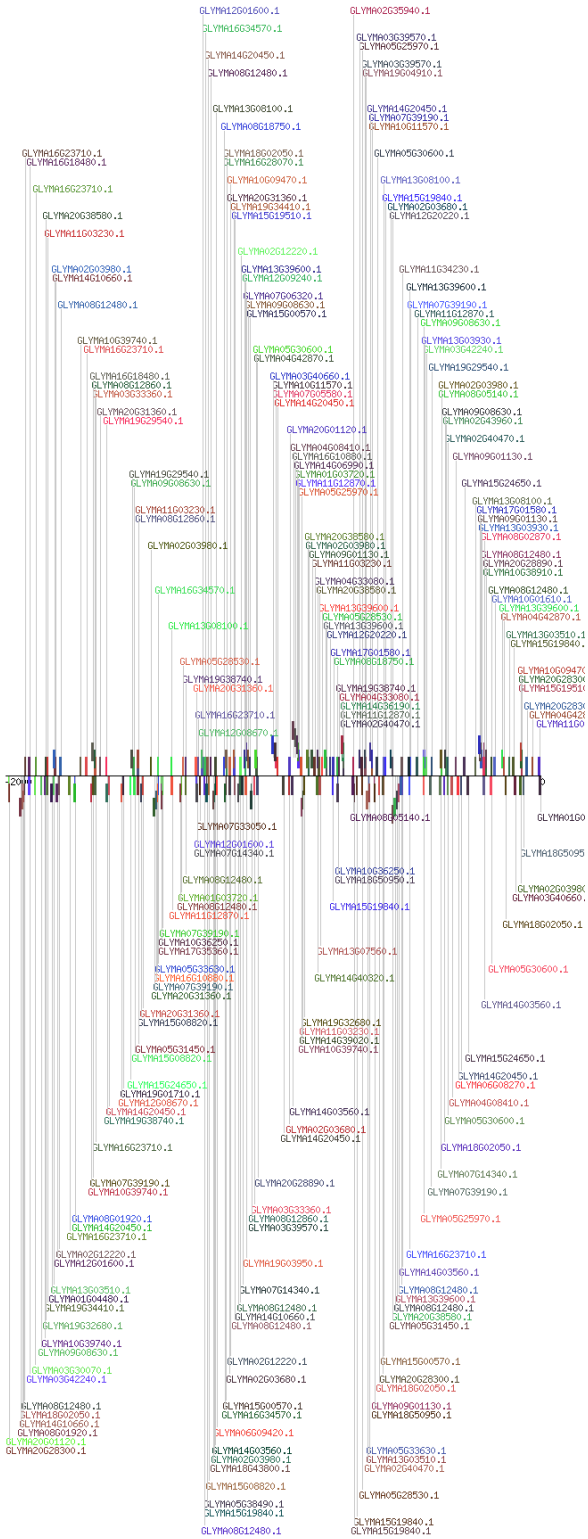


Figure 5. Distribution of the 2nd hit motif, DPBF 1& 2, in the promoters of the Leaf Preferential group (promoter length of 2kb). Figure was generated by Elefinder.

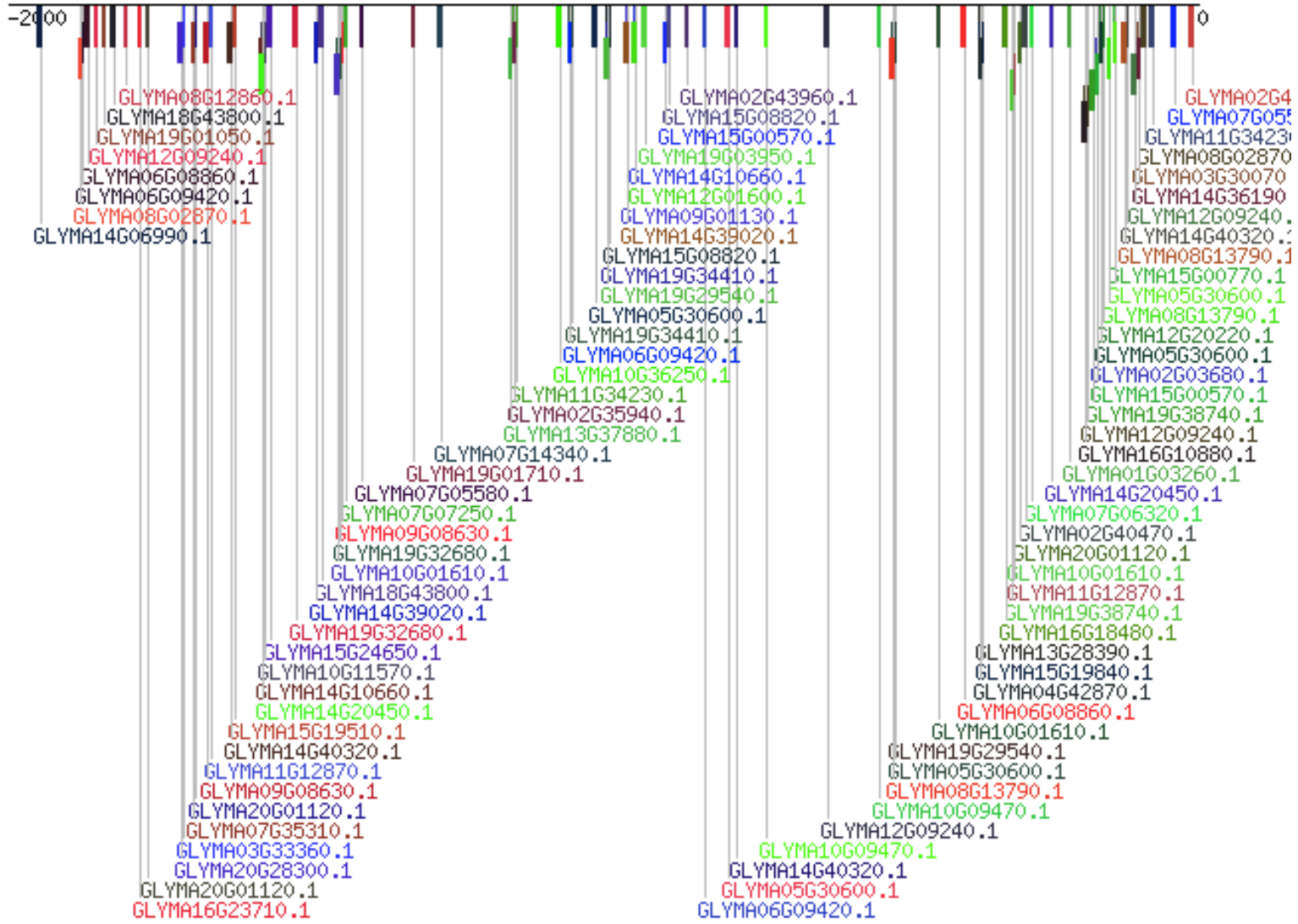


Figure 6. Distribution of the top hit motif, W- box, in the promoters of the Root Specific group (promoter length of 2kb). Figure was generated by Elefinder.

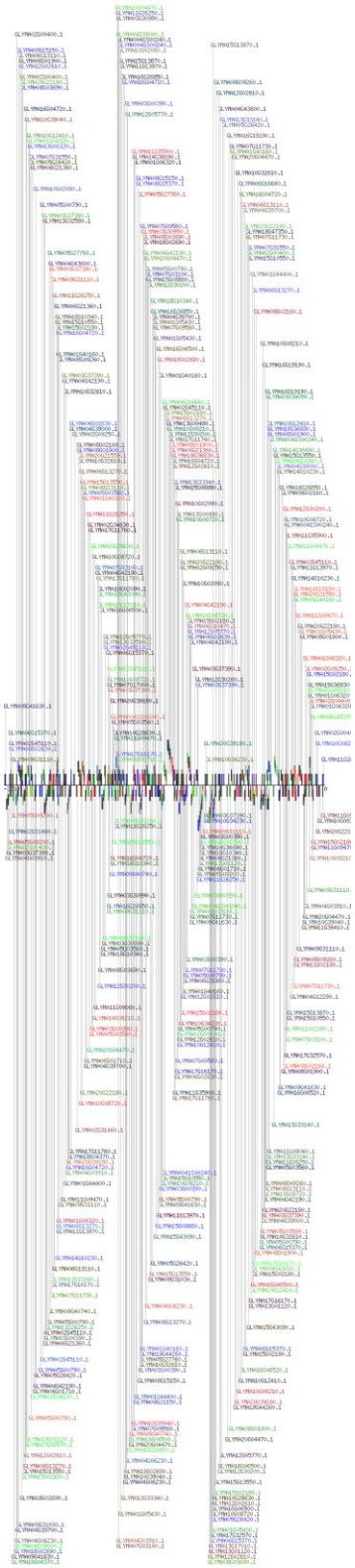


Figure 7. Distribution of the 2nd hit motif, Bellringer/replumless/pennywise BS1 IN AG, in the promoters of the Root-Specific group (promoter length of 2kb). Figure was generated by Elefinder.

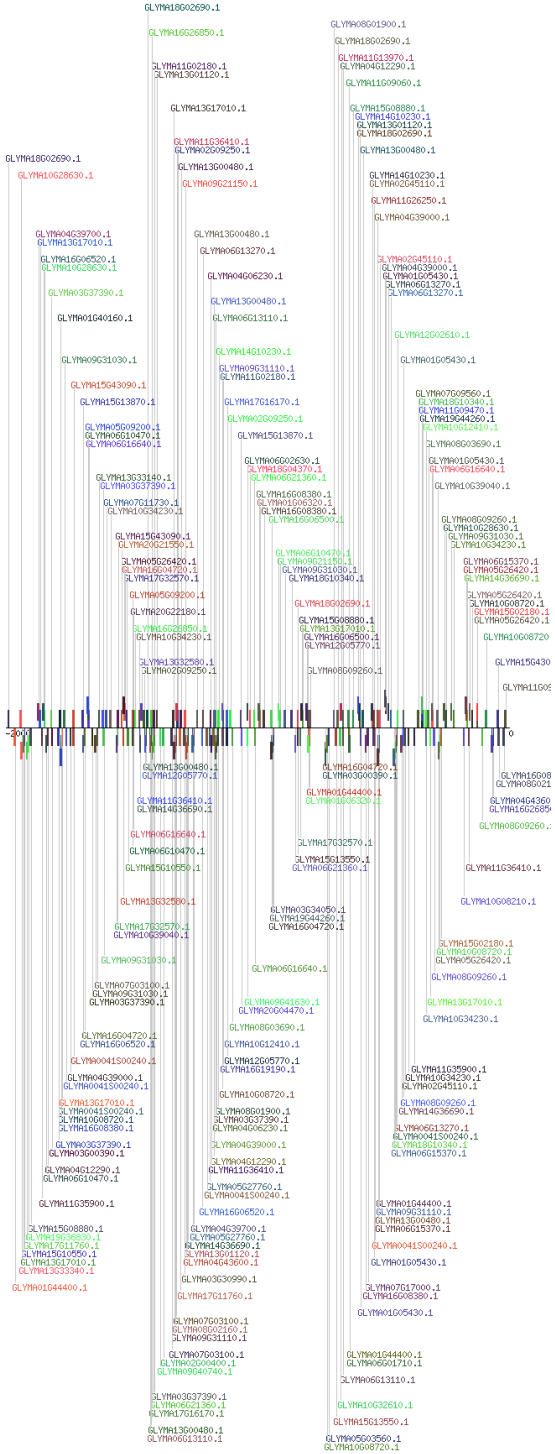


Figure 8. GO term enrichment for the 17 Constitutive genes found using REViGO. GO terms were found by using agriGO with the soybean genome locus as the reference background and *G. max* as the selected species. The agrioGO-generated list of GO terms was imported into REViGO for visualization. It was run with medium similarity allowed, and Arabidopsis was a reference of GO term sizes. The plot axes do not have an inherent meaning; the plot should be interpreted by the GO terms, how they cluster, their color (representing the $\log_{10}$p-value), and their size (larger size means a less specific GO term). The most significant GO term is translation for the constitutive gene set (p-value of 0.00014).

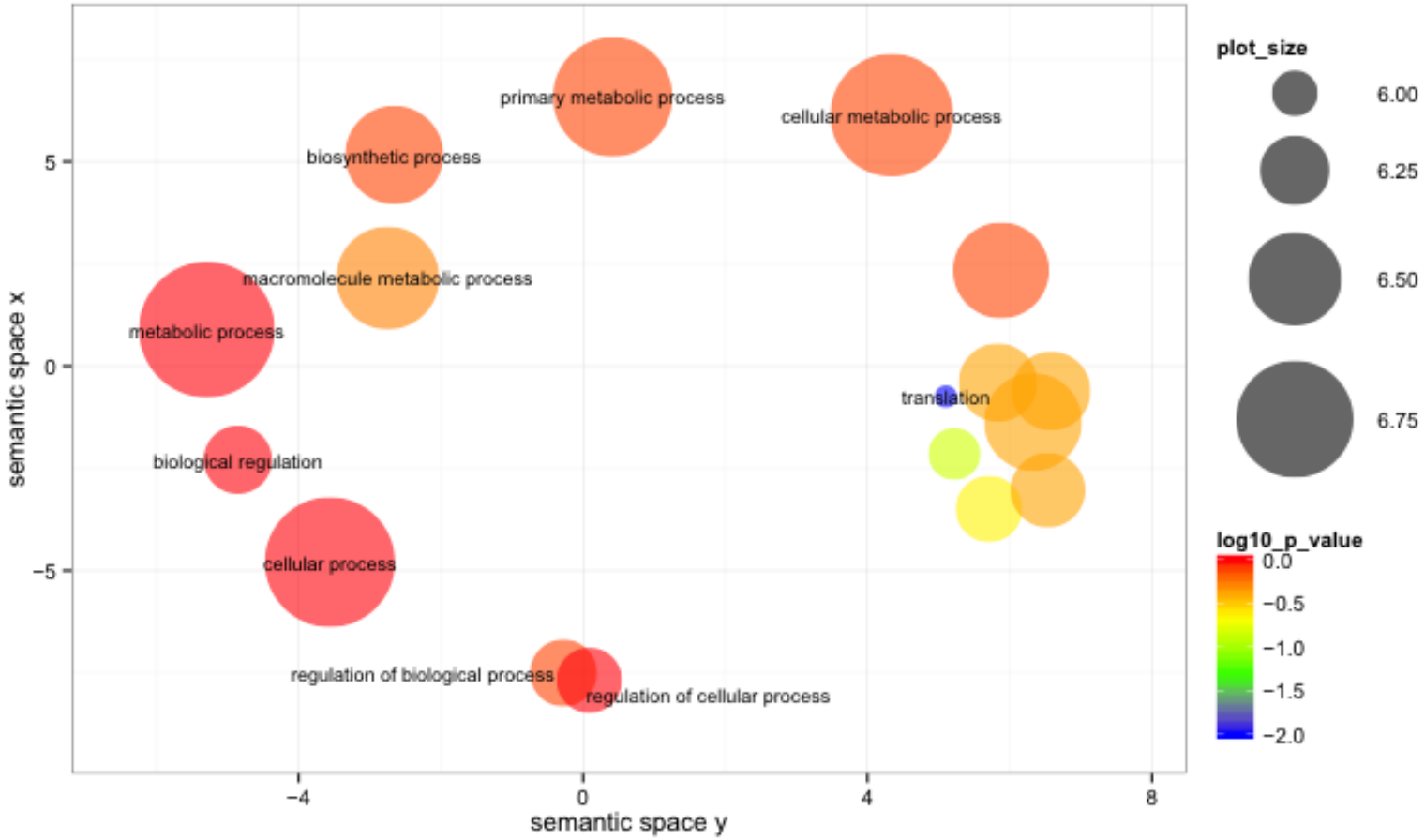


Figure 9. GO term enrichment for the top 100 Leaf-preferential genes found using REViGO. GO terms were found by using agriGOwith the soybean genome locus as the reference background and *G. max* as the selected species. The agrioGO-generated list of GO terms was imported into REViGO for visualization. It was run with medium similarity allowed, and Arabidopsis was a reference of GO term sizes. The plot axes do not have an inherent meaning; the plot should be interpreted by the GO terms, how they cluster, their color (representing the $\log_{10}$p-value), and their size (larger size means a less specific GO term). The most significant GO terms for the leaf-preferential set are photosynthesis and photosynthesis, light reaction (p-values of 2.4e-12 and 2.2e-05, respectively).

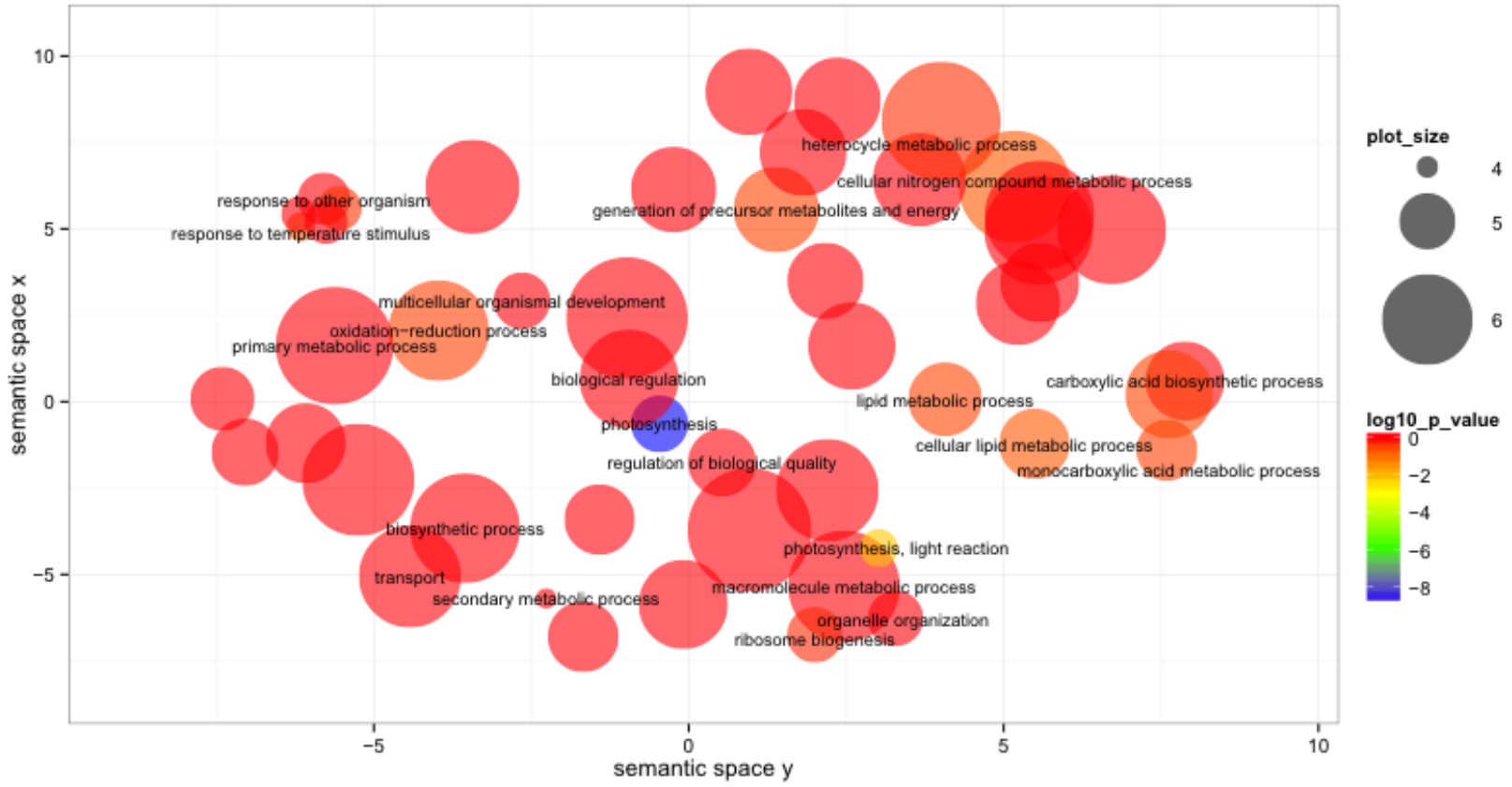
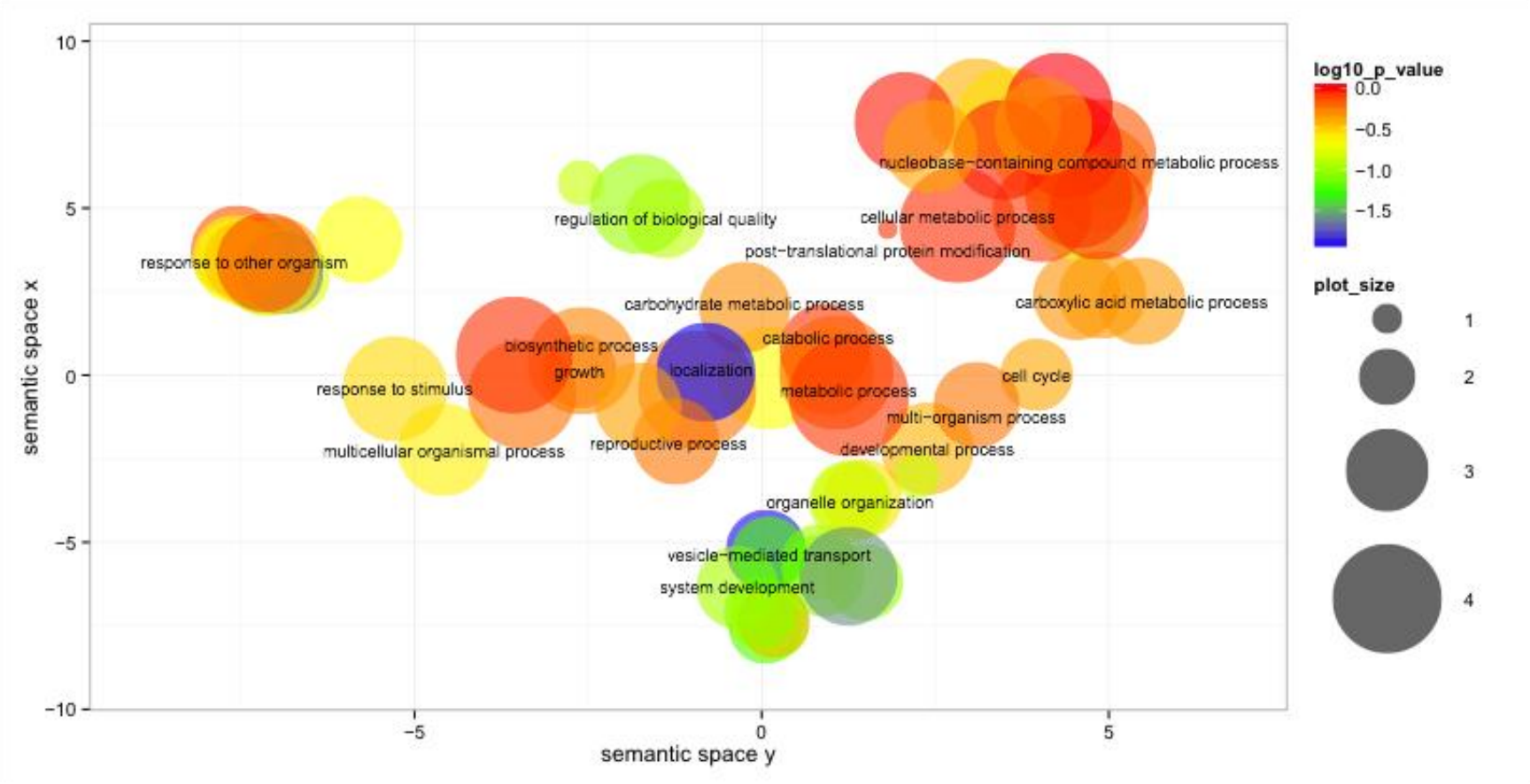


Figure 10. GO term enrichment for the top 100 root-specific genes found using REViGO. GO terms were found by using agriGO with the soybean genome locus as the reference background and *G. max* as the selected species. The agrioGO-generated list of GO terms was imported into REViGO for visualization. It was run with medium similarity allowed, and Arabidopsis was a reference of GO term sizes. The plot axes do not have an inherent meaning; the plot should be interpreted by the GO terms, how they cluster, their color (representing the $\log_{10}$p-value), and their size (larger size means a less specific GO term). Among the most significant GO terms for the root-specific set are localization and vesicle-mediated transport (p-values of 0.011 and 0.058, respectively).



CHAPTER III METHODS

Data acquisition and preliminary processing

To find strong promoters that confer the expression patterns of interest (constitutive, leaf-preferential, or root-specific), a data set with a wide representation of soybean tissues was needed. EST library data was downloaded from the NCBI UniGene database for *Glycine max*, build #39 (http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=3847). Transcripts within this database are placed into clusters that represent one gene, called a UniGene.

EST libraries were used for analysis only if they could give an accurate representation of a single soybean tissue's transcriptome. Thus, libraries that were normalized, made from more than 1 tissue, made from a tissue under stress inducing conditions, had less than 1000 ESTs or were a re-rack of a previous library were removed. After this filtering, 32 libraries were eligible for analysis. Table 1 lists these libraries. Further, transcript counts (for each individual Unigene) within a library were converted to percent abundance figures. Data within a library was sorted by decreasing percent abundance, since this study focuses on abundant transcripts.

The 32 EST libraries were grouped according to tissue and developmental stage, for ease in making comparisons between transcriptome profiles (Figure 1). The resulting groups represented 9 tissues: cotyledon, flower, leaf, pod, roots, seed coat, somatic embryos, stem, and germinating seeds. Table 2 shows the grouped libraries. The mean and standard error of the percent abundance term were calculated for unigene entries that occurred more than once in a group of EST libraries. No standard error cutoff was made.

Transcript functional annotations using soybean, *Arabidopsis* databases

Consensus sequences are not provided for individual unigenes; however, sequences with the longest region of high quality sequence data is provided. Thus, the representative Unigene sequences were annotated by using BLAT (set at 97% minimum identity) against the soybean longest high confidence coding sequence (CDS) (Kent, 2002; Schmutz et al., 2010). BLAT was chosen because it was designed to work on short sequences and can be

used to map ESTs to a genome (Kent, 2002; Nagaraj et al., 2006). The top hit for each Unigene was used as its soybean transcript identifier. Unigenes without a *G. max* hit were removed. The soybean CDS set was then run against the Arabidopsis Information Resource (TAIR)9cDNA set using TBLASTX with an e-value cutoff of 1e-4 (www.arabidopsis.org; Huala et al., 2001; Altschul et al., 1990). Entries that did not contain an *Arabidopsis* hit were removed. Since the *Arabidopsis* annotations are more reliable than *G. max* annotations, *Arabidopsis* orthologs were added. Further, these functional annotations were beneficial in determining if the pipeline to identify potential constitutive, leaf-preferential, or root-specific promoters was successful.

Identification of transcripts of interest

To find transcripts that have expression patterns of interest (constitutive, leaf-preferential, root-specific), the grouped EST libraries were selectively filtered. To identify strong promoters that confer high expression, all transcripts with percent abundances of less than 0.023 were disqualified. This is because Ubiquitin 10, a known highly expressed transcript has an average percent abundance level of 0.023 in this data set. This requirement ensures that strong promoters will be more likely to be identified.

To identify constitutive promoters, constitutive unigenes were first found. A Perl script was run to isolate the unigenes that were expressed in all grouped libraries above 0.023%. Likewise, a Perl script was used to filter the grouped library data to identify leaf-preferential unigenes. Unigenes were considered leaf-preferential if they expressed at their highest level in the leaf tissue group while being absent in the seed or seed coat groups. Finally, another Perl script was used to identify unigenes that occur only in the root group and defined as root-specific unigenes.

Functional annotation visualization

Gene functions for each group (constitutive, leaf-preferential, root-specific) were explored with gene ontology (GO) enrichment tools. GO terms were assigned to the 17 constitutive transcripts, the top 100 leaf-preferential transcripts, and the top 100 root-specific transcripts using agriGO (bioingo.cau.edu.cn/agriGO; Zhou et al., 2010). AgriGO analysis

was performed with the soybean genome locus as the reference background and *G. max* as the selected species. AgriGO returned lists of GO terms that were enriched in the transcripts of each group. Each list was imported into REViGO for visualization (revigo.irb.hr; Supek et al, 2011). REViGO was run with medium similarity allowed, and Arabidopsis was a reference of GO term sizes. For each of the 3 categories, REViGO generated an R script. Using these scripts, the plots were re-done locally in R.

Promoter identification and extraction

For the constitutive, leaf-preferential and root-specific unigenes that passed the cutoff criteria, the promoters of their corresponding soybean genes were extracted (found using BLAT). A gene's promoter was defined as the region immediately upstream of the annotated transcription start site (Figure 2). Promoters were obtained using the Soybean Genome Project Glyma 1.0 data (from Phytozome version 6) and an in-house promoter extraction script (phytozome.org; Schmutz et al., 2010). Promoters of minimal length (500 bp) and a longer, less conservative length (2kb) were extracted using this method.

Cis-regulatory promoter motif discovery

To identify promoter motifs for the promoters of each of the three expression categories, Sift and Elefinder were used (Hudson and Quail, 2003). Both tools can be accessed at http://stan.cropsci.uiuc.edu/tools.php. Sift was used to find novel motifs and Elefinder was used to find previously characterized motifs. Both Sift and Elefinder are used to find promoter motifs that are over-represented in a co-regulated gene set as compared to a reference gene set. The underlying assumption is that genes within a category (e.g. constitutive, leaf-preferential, and root-specific) have the same expression pattern because they are regulated by a specific transcription factor.

Sift was previously developed for novel motif detection in *Arabidopsis* Affymetrix microarray data (Hudson and Quail, 2003). Here, Sift was modified to detect novel motifs of 6 to 9 nucleotides in soybean, using a database generated from Soybean Genome Project Glyma 1.0 data (phytozome.org; Schmutz et al., 2010). The database is composed of the frequency of all possible enumerated motifs (from length 6 to 9) over all promoters in the

soybean genome. Sift was run on the longer promoters (2kb) for each of the constitutive, leaf-preferential and root-specific groups. We verified the P-values generated by Sift with the hypergeometric method. False discovery rate (FDR) was controlled at <5% to correct for multiple testing error (Benjamini and Hochberg, 1995).

Elefinder was used to detect previously characterized plant promoter motifs for each of the constitutive, leaf-preferential and root-specific groups. Elefinder has been updated to discover previously characterized motifs in nineteen plant species. This new multi-species Elefinder can discover motifs for the following plant species: *Aquilegia coerulea*, *Arabidopsis thaliana*, *Brachypodium distachyon, Chlamydomonas reinhardtii, Citrus clementine, Eucalyptus grandis, Glycine max, Medicago truncatula, Mimulus guttatus, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Prunus persica, Selaginella moellendorffii, Setaria italic, Sorghum bicolor, Vitis vinifera, Volvox carteri,* and *Zea mays*. As with Sift's database, the database for Elefinder is composed of motifs and their frequency in the promoters of the genome of interest. However, Elefinder's database is composed of motifs that are already characterized. These motifs were obtained from the Arabidopsis Gene Regulatory Information Server (AGRIS) Arabidopsis *cis*-regulatory element database (AtcisDB) (http://arabidopsis.med.ohio-state.edu; Davuluri et al., 2003).

Elefinder was run with *G. max* selected on the promoters of minimal length (500 bp) and a less conservative length (2kb) to discover over-represented motifs. P-values were calculated with the hypergeometric distribution, and false discovery rate was controlled at <5% (Benjamini and Hochberg, 1995).

CHAPTER IV DISCUSSION

In this work, transcriptome data representing individual tissues has been used to discover tissue-specific, tissue-preferential and constitutive promoters. In our pipeline, we used 32 single-tissue EST libraries, grouped them into 9 main tissues, and filtered for high-expressing transcripts. We extracted the promoters for these transcripts of interest and ran motif discovery tools on these sequence sets.

The promoter-discovery approach that we used requires a large dataset of transcript sequences from different tissues. We used publicly available data, which has many drawbacks. Although the EST libraries used were mainly from the Soybean EST Project, many different researchers constructed them. Thus, the quality of this data can vary. Further, many different soybean varieties were used in this EST sequencing effort. The researchers of the Soybean EST Project noted in their publication that generally libraries made of the same tissue usually clustered together (as represented in a dendogram made by clustering on contig content similarity) (Shoemaker et al., 2002). However, some single-tissue expression libraries did not cluster with libraries made from the same tissue from different varieties, showing that not all soybean varieties showed the same expression profile for a specific tissue (Shoemaker et al., 2002). Another possibility is that subtle variation in the plant culture conditions used by different researchers caused a different subset of transcripts to be expressed. Further, we used the soybean reference sequence (*var.* Williams 82) in our annotation efforts. This genome does not represent all of the genome variation for *Glycine max*. Naturally, any promoters or motifs found in this study should be experimentally verified. Transgenic *Arabidopsis* promoter:Citrine fusion constructs are in the process of being made for the strongest 3 constitutive, 2 root-specific and 2 leaf-preferential (not in seed or seed coat) promoters. This experimental design is not ideal; we assume that soybean promoters will confer expression in *Arabidopsis* in the same manner as *in vivo.* Soybean transgenics are laborious and time-consuming to create; thus, we chose to use *Arabidopsis* for our transgenic system.

As previously noted, annotation for clustered transcripts (unigenes) as provided from NCBI was not sufficient for our analysis. Each soybean unigene cluster was mapped to the soybean genome in a conservative manner. The database used, the *G*. *max* high confidence CDS, included only genes that were most likely to be functional (Schmutz et al., 2010). The low confidence CDS file was not used for annotation. Further, the BLAT query (the unigene sequence file) contained the longest region of high quality sequence for each unigene present in *G. max* (Kent, 2002). It was not optimal that there was not a consensus sequence made for each unigene, or that the representative unigene sequence may not even be from an EST library used in this study. Further, when we performed this study, *G. max* did not contain an annotation file for the genome. We had to use *Arabidopsis* ortholog annotations, as identified by TBLASTX of the soybean coding sequence (CDS) file against the Arabidopsis Information Resource (TAIR)9 cDNA set (Altschul et al., 1990). *Arabidopsis* and soybean are highly diverged species, but *Arabidopsis* was the closest organism with reliable functional annotations.

The most abundant transcripts for each of the three desired expression patterns (constitutive, leaf-preferential but not in seed or seed coat, and root-specific) are functional annotations that are consistent with the processes that occur in such tissues. The most abundant constitutive transcripts are Glyma13g17830.1 and Glyma11g37970.1, which are annotated as ubiquitin 14 and histone H3, respectively. Histones are involved in DNA packaging, which all cells must perform. Further, ubiquitins are involved in protein degradation, which is a normal cellular process. The third most abundant constitutive transcript is Glyma07g39020.1 (PRXR1, a peroxidase). In *Arabidopsis*, this transcript is known to be expressed at extremely high levels, specifically 24.5 times more than the average expression level of other genes (Thierry-Mieg and Thierry-Mieg, 2006). The most highly expressed constitutive transcripts identified are consistent with those that are previously known to be globally and highly expressed. Likewise, the GO term enrichment analysis revealed “transcription” as the most significant term for the promoters of this group. All tissues must transcribe RNAs to function properly, so this designation is intuitive.

The highly expressed transcripts identified to be leaf-preferential (not expressed in seed or seed coat) are annotated to be mainly photosynthesis-related. The most abundant two transcripts are Glyma11g34230.1 (Rubisco activase) and Glyma10g39740.1 (THI1, thiazole biosynthetic enzyme). Rubisco activase is known to be involved in photosynthesis, and the THI1 protein is targeted to chloroplasts (Ribeiro et al., 2005). However, the *Arabidopsis* THI1 promoter has been shown to not confer leaf-preferential expression. It has been demonstrated that THI1 is constitutively expressed, and highly expressed in shoots and to a lesser degree in roots of mature plants (Ribeiro et al., 2005). This may indicate that the previously characterized expression pattern could be *Arabidopsis*-specific, it could indicate problems with the *Arabidopsis* study, or it could show a deficiency in our promoter discovery method. The desired expression pattern for this group is highly specific (leaf-preferential not in seed or seed coat), which our method could not reliably detect.

GO term enrichment analysis for the leaf-preferential group detected many photosynthetic-related terms. The most significant terms were "photosynthesis" and "photosynthesis light reaction" (p-values of 2.4e-12 and 2.2e-05, respectively). It would be expected that leaf-preferential transcripts should be enriched in photosynthetic functions, for that is a main function of the leaf tissue. However, many other GO terms were found that were not significant (has low p-value, high –log10 value, and is red in graph) and did not have much functional overlap (shown by clustering in graph). This group of non-significant unrelated GO terms could be due to the fact that this group is not leaf-specific, which would reduce the number of GO terms. However these terms may just represent noise in the data.

The root-specific group contained many annotated ion binding and localization-related transcripts. The top two transcripts were Glyma12g05770.1 and Glyma15g13550.1 (Beta Glucosidase 17, BGLU17, and peroxidase 22, PER22, respectively). The proteins of these two genes are both proposed to localize in the endomembrane system (Thierry-Mieg and Thierry-Mieg, 2006). BGLU17 has been proposed to act as a hydrolase, while the PER22 has been proposed to be involved in metabolism (Thierry-Mieg and Thierry-Mieg, 2006). GO term enrichment for the top 100 genes of this group found localization and vesicle-mediated transport as the most significant terms. Further, many GO terms were involved in

metabolism and response to stimuli or other organisms. The enriched terms reflect the function of the roots in nutrient uptake and symbioses.

Sift and Elefinder were used to discover over-represented motifs in promoters for each group. Sift was used to find novel motifs, Elefinder for known motifs. The underlying assumption for these tools is that for co-regulated genes, a common transcription factor (or set of factors) will control their expression. Transcription factor binding cannot be reliably modeled *in silico*; thus, we rely on techniques to find the conserved sites (motifs) that these proteins bind to. Elefinder was more likely to find motifs that are overrepresented, because it does a smaller number of tests. Sift performs millions of tests, because it considers all possible short, enumerated words (representing motifs). When correcting for false discovery rate (FDR), many potential significant motifs are discarded due to this large test number. Considering Sift's low power due to multi-testing error, it was run only with longer promoters (2kb). Motif discovery tools run with minimal promoters may give more consistent results, but longer promoters are more sensitive. Further, 2kb promoters can compensate for incorrectly annotated transcription start sites. Elefinder was run on both minimal (500bp) and larger promoters (2kb).

No significant results were found for both the leaf-preferential and the constitutive 2kb promoters using Sift. "Significance" here refers to motifs that were able to pass the FDR cutoff of <5%. The top novel motif hit for the constitutive promoter set, CGTCGNTT, was able to pass the FDR cutoff if it was increased to <10%. The root-specific promoter set detected a novel promoter motif, GNAATNTCA, which was significant at an FDR cutoff of <5%.

Elefinder was more successful at finding significant motifs for promoters of the three groups. Elefinder was first run on minimal promoters, a more conservative test. The motifs detected for the constitutive promoters and the root-specific promoters were not significant at an FDR cutoff of <5%. No significant motifs were detected with the minimal leaf-preferential promoters. We were unable to find any significant motifs in the core promoters using Elefinder.

More motifs were detected using Elefinder with longer promoters. The constitutive promoters did not yield any significant results. Perhaps the promoter set used is not large enough (only 17 promoters were in the analysis set), or there is not a common set of transcription factors that regulate them. The top 100 leaf-preferential promoters had two significantly over-represented motifs, Bellringer (AAATTAAA) and DPBF1&2 (ACACNNG) (Bao et al., 2004; Kim, et al., 1997). The DPBF1&2 (*Dc3* Promoter-Binding Factor 1&2) motif was characterized in the promoter of a gene strongly expressed in plant embryogenesis (Kim, et al., 1997). The top 100 root-specific promoters had three significant over-represented motifs, W-box (TTGAC), Bellringer (AAATTAAA), and RAV1-A (CAACA) (Yu et al., 2001; Kim, et al., 1997; Kagaya et al., 1999). WRKY transcription factors, which are implicated in defense responses, bind W-box motif motifs (Yu et al., 2001). The transcription factor RAV1, which may be involved in plant growth and development, binds to the RAV1-A motif (Hu et al., 2004). It is interesting that both the root-specific and leaf-preferential promoters shared a significantly over-represented motif, Bellringer. It is possible that transcription factors that bind this motif are part of a previously uncharacterized pathway.

Considering the small number of statistically over-represented motifs found using Sift and Elefinder, other motif–finding tools could be harnessed. For example, promoter motif tools that use phylogeny decrease the rate of false positive motifs found. When this study was being performed, there were not any close relatives of soybean that had reference sequences. Recently, the genome for *Phaseolus vulgaris*, a close relative of soybean, has become available. As more plant genomes are completed, motif-discovery tools that incorporate phylogeny could be used to discover binding sites in soybean promoters.

CHAPTER V CONCLUSION

In this study, we analyzed available transcriptome data for soybean to discover strong tissue-specific, tissue-preferential and constitutive promoters. First, we found transcripts that had the expression patterns of interest (root-specific, leaf-preferential but not in seed or seed coat, and constitutive). Next, we extracted promoters for these abundant transcripts. We conclude that we have identified a set of probable constitutive, leaf-preferential, and root-specific promoters, based upon gene annotation and GO term analysis. Experimental verification for these top promoters is forthcoming.

After performing motif analysis on the sets of co-regulated genes (leaf-preferential, root-specific and constitutive), we have found several statistically over-represented motifs. A novel motif, GNAATNTCA, was found in the root-specific promoter group. Several over-represented known motifs were found for the leaf-preferential and the root-specific group using Sift. Unfortunately, motif characterization is a naïve view of regulation of gene expression. It only examines a part of transcriptional regulation and neglects to consider the effect of epigenetics (methylation, chromatin remodeling, etc.). Thus, our results do not show the whole picture of transcriptional regulation for these promoters of interest. However, it does add to the overall body of knowledge on promoter-regulated transcriptional regulation.

REFERENCES

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nat Rev Genet* **11**, 559–571 (2010).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J MolBiol* 215, 403–10 (1990).

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796-815

Bao, X., Franks, R. G., Levin, J. Z. & Liu, Z. Repression of AGAMOUS by BELLRINGER in Floral and Inflorescence Meristems. *The Plant Cell Online* **16**, 1478–1489 (2004).

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57:289-300

Brown, T. *Gene Cloning and DNA Analysis: An Introduction*. (Wiley-Blackwell: 2010).

Gibney, E. R. & Nolan, C. M. Epigenetics and gene expression. *Heredity* **105**, 4–13 (2010).

Hu YX, Wang YH, Liu XF, & Li JY (2004) Arabidopsis RAV1 is down-regulated by brassinosteroid and may act as a negative regulator during plant development. Cell Research (2004) 14, 8–15

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, HuangW, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. NucleicAcids Res. 29: 102-105

Hudson ME, Quail PH (2003) Identification of Promoter Motifs Involved in the Network of Phytochrome A-Regulated Gene Expression by Combined Analysis of Genomic Sequence and Microarray Data. Plant Physiol. 133: 1605-1616

Hudson, ME. Sequencing breakthroughs for genomic ecology and evolutionary biology. *MolEcolResour.* **8**, 3–17 (2008).

Jones, N. & Pevzner, P. *An Introduction to Bioinformatics Algorithms*. (MIT Press: Cambridge, MA, USA, 2004).

Kagaya, Y., Ohmiya, K. & Hattori, T. RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. *Nucleic Acids Research* **27**, 470–478 (1999).

Kaufmann, K., Pajoro, A. &Angenent, G. C. Regulation of transcription in plants: mechanisms controlling developmental switches. *Nat Rev Genet* **11**, 830–842 (2010).

Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Research* 12, 656–664 (2002).

Kim, S. Y., Chung, H.-J. & Thomas, T. L. Isolation of a novel class of bZIP transcription factors that interact with ABA-responsive and embryo-specification elements in the Dc3 promoter using a modified yeast one-hybrid system. *The Plant Journal* **11**, 1237–1251 (1997).

Mount, D. *Bioinformatics: Sequence and Genome Analysis*. (Cold Spring Harbor Laboratory Press: New York, 2001).

Nagaraj, S. H., Gasser, R. B. & Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*8, 6–21 (2007).

Pontius, J. U., Wagner, L. & Schuler, G. D.*The NCBI Handbook*. (2002).

Pontius, J. U., Wagner, L. & Schuler, G. D. UniGene: A Unified View of the Transcriptome. *The NCBI Handbook* (2002) at <http://www.ncbi.nlm.nih.gov/books/NBK21083/>

Ramana Davuluri, Hao Sun, Saranyan Palaniswamy, Nicole Matthews, Carlos Molina, Mike Kurtz, and Erich Grotewold. AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics, 4(1):25+, June 2003.

Ribeiro, D. T. *et al.* Functional characterization of the thi1 promoter region from Arabidopsis thaliana. *Journal of Experimental Botany* **56**, 1797–1804 (2005).

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. *Genome sequence of the palaeopolyploid soybean.* Nature. 2010 Jan 14; 463(7278):178-83.

Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW, Waterston R, Smoller D, Coryell V, Khanna A, Erpelding J, Gai X, Brendel V, Raph-Schmidt C, Shoop EG, Vielweber CJ, Schmatz M, Pape D, Bowers, Y, Theising B, Martin J, Dante M, Wylie T, Granger C. A compilation of soybean ESTs: generation and analysis. *Genome* 45, 329–338 (2002).

Siddharthan, R., Siggia, E. & Van Nimwegen, E. PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoSComputBiol* **1**, e67 (2005).

Singh, K. B. Transcriptional Regulation in Plants: The Importance of Combinatorial Control. *Plant Physiology* **118**, 1111–1120 (1998).

Sinha, S., Blanchette, M. & Tompa, M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5**, (2004).

Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).

Supek F, Bošnjak M, Škunca N, Šmuc T. "*REVIGO summarizes and visualizes long lists of Gene Ontology terms*" PLoS ONE 2011. doi:10.1371/journal.pone.0021800

Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biology **7**, S12 (2006).

Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotech* **23**, 137–144 (2005).

Walley, J. *et al.* Mechanical Stress Induces Biotic and Abiotic Stress Responses via a Novel cis-Element. *PLoS Genet* **3**, e172 (2007).

Wasserman, W. W. &Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276–287 (2004).

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pontius, J. U., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L., and Yaschenko, E. Database resources of the National Center for Biotechnology. *Nucleic Acids Research* 31, 28–33 (2003).

Yu, D., Chen, C. & Chen, Z. Evidence for an Important Role of WRKY DNA Binding Proteins in the Regulation of NPR1 Gene Expression. *The Plant Cell Online* **13**, 1527–1540 (2001).

Zhou Du, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su agriGO: a GO analysis toolkit for the agricultural community Nucleic Acids Research Advance Access published on July 1, 2010, DOI 10.1093/nar/gkq310. Nucl. Acids Res. 38: W64-W70.

APPENDIX A: CONSTITUTIVE TRANSCRIPTS THAT PASSED INITIAL SELECTIVE FILTERING

Using the promoter discovery pipeline, we found 48 constitutive transcripts that were both in every grouped library, and above the level of ubiquitin 10. These transcripts are listed in **SupplementalTable1.xlsx**. These transcripts underwent further filtering to remove transcripts that were not expressed at the same level in every grouped library (See results).

APPENDIX B: LEAF-PREFERENTIAL (ABSENT IN SEED OR SEED COAT) TRANSCRIPTS THAT PASSED SELECTIVE FILTERING

Leaf-preferential (absent in seed or seed coat) transcripts were defined as those that were expressed highest in leaf, but did not occur in the seed or seed coat. Also, transcripts were required to express higher than ubiquitin 10 to pass the filtering step.

SupplementalTable2.xlsx contains the 1401 leaf-preferential (absent in seed or seed coat) transcripts that passed filtering criteria.

APPENDIX C: ROOT-SPECIFIC TRANSCRIPTS THAT PASSED SELECTIVE FILTERING

Root-specific transcripts were defined as transcripts that occurred only in the root grouped library (and must be expressed at a level greater than ubiquitin 10). The 219 high-expressing root-specific transcripts are listed in **SupplementalTable3.xlsx.**

APPENDIX D: PROMOTER SEQUENCES FOR THE TOP EXPRESSED CONSTITUTIVE TRANSCRIPTS

After selectively filtering for constitutive transcripts, 17 such transcripts were identified. The corresponding genes for these transcripts were identified. Next, the promoters were extracted using BLAST command line resources, soybean genome information, and a Perl script. Constitutive promoter sequences (length of 2kb) are provided in **SupplementalFile1.txt**.

APPENDIX E: PROMOTER SEQUENCES FOR THE TOP 100 EXPRESSED LEAF-PREFERENTIAL (ABSENT IN SEED OR SEED COAT) TRANSCRIPTS

After applying all filtering criteria for leaf-preferential (absent in seed or seed coat) transcripts, 1,401 transcripts were identified. The promoters of the top 100 expressed transcripts were extracted using a Perl script, soybean genome information, and BLAST command line resources. The promoter sequences (length of 2kb) for the top 100 transcripts are contained in **SupplementalFile2.txt**.

APPENDIX F: PROMOTER SEQUENCES FOR THE TOP 100 EXPRESSED ROOT-SPECIFIC TRANSCRIPTS

201 transcripts were identified after selectively filtering for root-specific transcripts. Promoters for the top 100 expressed root-specific transcripts were extracted. An in-house Perl script, BLAST command line resources, and soybean genome information were used for promoter extraction. Promoter sequences (length of 2kb) for these top-expressed transcripts are included in **SupplementalFile3.txt**.